

Simulation-Based Hypothesis Testing of High Dimensional Means Under Covariance Heterogeneity — An Alternative Road to High Dimensional Tests

Jinyuan Chang[†], Wen Zhou[‡], Wen-Xin Zhou[§]

April 20, 2015

Abstract

In this paper, we study testing the population mean vector of high dimensional multivariate data for both one-sample and two-sample problems. The proposed simulation-based testing procedures employ maximum-type statistics and use the Gaussian approximation techniques to obtain corresponding critical values. Different from peer tests that heavily rely on the structural conditions on the unknown covariance matrices, the proposed tests allow very general forms of the covariance structures of data and therefore enjoy wide scope of applicability in practice. To enhance powers of the tests against sparse alternatives, we further propose two-step procedures with a preliminary feature screening step. Theoretical properties of the proposed tests are investigated. Extensive numerical experiments on synthetic datasets and empirical applications on identifying diseases-associated gene-sets are provided to support the theoretical results. The proposed tests are easily implemented and computationally efficient in practice.

Keywords: Feature screening; Gaussian approximation; High dimensional hypothesis test; Simulation-based statistical inference; Testing equality of mean vectors

1 Introduction

High dimensional hypothesis test for mean vectors is in great need in multiple scientific disciplines and practical fields, particularly in modern genomics, image segmentation and quantitative

[†]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia.
E-mail: jinyuan.chang@unimelb.edu.au.

[‡]Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.
E-mail: riczw@stat.colostate.edu.

[§]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia.
E-mail: wenxin.zhou@unimelb.edu.au.

finance. With limited number of subjects being measured, the sample size is relatively small which leads to the so-called “large p , small n ” problem. Furthermore, the measurements usually possess fairly complex dependence structures. These issues have urged the renovation of traditional multivariate analysis procedures on testing mean vectors. See, for example, [Bai and Saranadasa \(1996\)](#), [Donoho and Jin \(2004\)](#), [Chen and Qin \(2010\)](#) and [Cai, Liu and Xia \(2014\)](#), among others.

In this paper, we consider both the one-sample and two-sample settings. Let \mathbf{X} and \mathbf{Y} be two p -dimensional random vectors with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Consider two independent random samples of independent and identically distributed observations, $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ and $\mathcal{Y}_m = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$, drawn from the distributions of \mathbf{X} and \mathbf{Y} , respectively. It is of general interest in testing the hypotheses

(i) (One-sample problem)

$$H_0^{(I)} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 \text{ versus } H_1^{(I)} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$$

for a *specified* p -dimensional vector $\boldsymbol{\mu}_0$, which, without loss of generality, is equivalent to

$$H_0^{(I)} : \boldsymbol{\mu}_1 = \mathbf{0} \text{ versus } H_1^{(I)} : \boldsymbol{\mu}_1 \neq \mathbf{0}; \quad (1.1)$$

(ii) (Two-sample problem)

$$H_0^{(II)} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ versus } H_1^{(II)} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \quad (1.2)$$

Hypothesis (1.1) arises when a sample is being compared to a hypothetical population with known prior information, while (1.2) sparks interest to compare two parallel groups, particularly a control group and a treatment group in scientific experiments. In the low dimensional setting where p is fixed, traditional tests have been extensively studied for testing both (1.1) and (1.2). For example, the properties for both the one-sample and two-sample Hotelling’s T^2 tests have been examined under the assumption of normality ([Anderson, 2003](#)). We refer to [Dembo and Shao \(2006\)](#) and [Liu and Shao \(2013\)](#) for the large and moderate deviation asymptotics for the Hotelling’s T^2 -statistics without normality.

Generally, the sum of squares-type and the maximum-type statistics are used to test the hypotheses (1.1) and (1.2) in the high dimensional settings. The sum of squares-type statistics aim to mimic the weighted Euclidean norms, $\|\mathbf{A}^{1/2}\boldsymbol{\mu}_1\|_2^2$ or $\|\mathbf{A}^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2^2$ for certain linear transformation \mathbf{A} , and the corresponding tests are powerful for detecting relatively dense signals ([Bai and Saranadasa, 1996](#); [Srivastava and Du, 2008](#); [Srivastava, 2009](#); [Chen and Qin, 2010](#)). Statistics of the maximum-type, on the other hand, are preferable for detecting relatively sparse signals ([Liu and Shao, 2013](#); [Cai, Liu and Xia, 2014](#)) and have been used in a variety of applications including the medical image problem ([James, Clymer and Schmalbrock, 2001](#)), anomaly detections

(Castagna, Sun and Siegfried, 2003) and gene selections (Martens et al., 2005).

Most existing testing procedures for (1.1) and (1.2) rely on the derivation the pivotal limiting distribution of test statistics, from which the critical value is approximated. In the high dimensional scenarios, various structural assumptions on the unknown covariance matrices have been imposed. For example, the thresholding test statistic proposed by Zhong, Chen and Xu (2013) for the one-sample problem (1.1) requires weak dependence to guarantee the asymptotic normality, that is, the observation is generated by the model $\mathbf{X} = \mathbf{W} + \boldsymbol{\mu}$ with weakly stationary sequence $\mathbf{W} = (W_1, \dots, W_p)'$ satisfying $\sum_{k \geq 1} |\text{Cov}(W_1, W_{k+1})| < \infty$. For the two-sample problem (1.2), Cai, Liu and Xia (2014) proposed a testing procedure whose validity requires that, among certain moment conditions, the unknown covariance matrices satisfy $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $\max_{1 \leq k < \ell \leq p} |\omega_{k\ell} / \sqrt{\omega_{kk}\omega_{\ell\ell}}| \leq c$ for precision matrix $\boldsymbol{\Omega} = (\omega_{k\ell})_{1 \leq k, \ell \leq p}$ and some $c \in (0, 1)$.

However, in many applications, these assumptions can be very restrictive or difficult to be verified, and therefore limit the scope of applicability for the limiting distribution calibration approach. First, the existence of a pivotal asymptotic distribution relies heavily on the structural assumptions on the unknown covariance/correlation structures, which may not be true in practice. For example, it is very common that the expression levels are highly correlated for genes regulated by the same pathway (Wolen and Miles, 2012) or associated with the same functionality (Katsani et al., 2014), which results in a complex and non-sparse covariance structure. These empirical evidences indicate that the strong structural assumptions on the covariance matrices may sometimes be unrealistic in real-world applications. Another concern, as pointed out by Cai, Liu and Xia (2014), is that the convergence rate to the extreme value distribution of maximum-type statistics is usually slow. Taking the extreme distribution of type I as an example, it was shown in Liu, Lin and Shao (2008) that the convergence rate is of order $O\{\log(\log n) / \log(n)\}$. Although the convergence rate may be improved by using suitable intermediate approximations, still its validity relies on the dependence structure of the underlying distribution.

Driven by the above two concerns, we revisit the problem of testing hypotheses (1.1) and (1.2) in high dimensions from a different perspective. Motivated by applications in genomic analysis and image analysis, we are particularly interested in detecting discrepancies when $\boldsymbol{\mu}_1$ and $\mathbf{0}$ or $\boldsymbol{\mu}_2$ are distinguishable to a certain extent in at least one coordinate. We develop a fully data driven procedure to compute the critical values using the Monte Carlo simulations under appropriate settings. The validity of our procedure is established without enforcing structural assumptions of any kind on the unknown covariances. The main idea is based on the strong approximation of empirical processes by Gaussian processes (van der Vaart and Wellner, 1996; Chernozhukov, Chetverikov and Kato, 2013), and to some degree, is similar to that of Liu and Shao (2013) that utilizes the intermediate approximation. However, instead of generating independent standard multivariate normal vectors, our approach takes into account correlations among the features and therefore is automatically adapted to the underlying dependence.

The rest of the paper is organized as follows. In Section 2, we describe the simulation-based testing procedures for both hypotheses (1.1) and (1.2). Theoretical properties of the tests are studied in Section 3. Numerical studies are reported in Section 4 to assess the performance of the proposed tests comparing to the peer methods. In Section 5, we applied the proposed tests to an acute lymphoblastic leukemia data for identifying disease-associated gene-sets based on the gene expression levels. The paper is concluded with a brief discussion in Section 6. The underpinning technical details, as well as additional simulation results and empirical data analysis, are relegated to the supplementary material.

2 Methodology

We present our testing procedures for the one-sample and two-sample problems in Sections 2.1.1 and 2.1.2, respectively. Section 2.2 discusses the connections between the proposed tests and covariance estimation, the latter of which is of independent interest. In Section 2.3, we develop a preliminary screening procedure to improve the power of proposed tests against sparse alternatives.

Throughout the paper, we denote by $|\boldsymbol{\beta}|_\infty = \max_{1 \leq k \leq p} |\beta_k|$ for a p -dimensional vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$. For a matrix $\mathbf{W} = (w_{k\ell})_{p \times p}$, its max-norm is defined as $\|\mathbf{W}\|_\infty = \max_{1 \leq k, \ell \leq p} |w_{k\ell}|$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there exist positive constants c_1, c_2 such that $c_1 \leq a_n/b_n \leq c_2$ for all $n \geq 1$, we write $a_n = O(b_n)$ if there is a constant C such that $|a_n| \leq C|b_n|$ for all sufficiently large n , and we write $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$. Moreover, let \mathbf{X} and \mathbf{Y} be two p -dimensional random vectors with means $\boldsymbol{\mu}_1 = (\mu_{11}, \dots, \mu_{1p})'$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \dots, \mu_{2p})'$, covariance matrices $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1 \leq k, \ell \leq p}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{2,k\ell})_{1 \leq k, \ell \leq p}$, respectively. Denote by \mathbf{R}_1 and \mathbf{R}_2 the corresponding correlation matrices. Let $\mathbf{D}_1 = \text{diag}(\boldsymbol{\Sigma}_1)$ and $\mathbf{D}_2 = \text{diag}(\boldsymbol{\Sigma}_2)$ be the diagonal matrices of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Let $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ and $\mathcal{Y}_m = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$ be two independent samples consisting of independent and identically distributed observations drawn from the distributions of \mathbf{X} and \mathbf{Y} , respectively. For each $i = 1, \dots, n$ and $j = 1, \dots, m$, write $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$ and $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{jp})'$. Set $N = n + m$.

2.1 Test procedures

2.1.1 One-sample case

Consider the maximum-type statistics in the following forms:

$$T_{\text{ns}}^{(\text{I})} = \max_{1 \leq k \leq p} \sqrt{n} |\bar{X}_k| \quad \text{or} \quad T_s^{(\text{I})} = \max_{1 \leq k \leq p} \frac{\sqrt{n} |\bar{X}_k|}{\hat{\sigma}_{1k}}, \quad (2.1)$$

where $\bar{X}_k = n^{-1} \sum_{i=1}^n X_{ik}$ and $\hat{\sigma}_{1k}^2 = n^{-1} \sum_{i=1}^n (X_{ik} - \bar{X}_k)^2$. Throughout, the statistic $T_s^{(\text{I})}$ is referred as the *studentized* statistic, while $T_{\text{ns}}^{(\text{II})}$ is referred as the *non-studentized* statistic. Intu-

itively, large values of $T_{\text{ns}}^{(1)}$ or $T_s^{(1)}$ provide evidences against $H_0^{(1)}$ in (1.1) so that the corresponding tests are of the form

$$\Psi_{\text{ns},\alpha}^{(1)} = I\{T_{\text{ns}}^{(1)} > \text{cv}_{\text{ns},\alpha}^{(1)}\} \quad \text{or} \quad \Psi_{s,\alpha}^{(1)} = I\{T_s^{(1)} > \text{cv}_{s,\alpha}^{(1)}\},$$

where $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$ are the critical/cut-off values. With $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$ properly chosen, we expect the above two tests to have, at least approximately, a prescribed size $\alpha \in (0, 1)$.

For $\nu \in \{\text{ns}, s\}$, $\text{cv}_{\nu,\alpha}^{(1)}$ is typically specified as the $(1 - \alpha)$ -quantile of the limiting distribution of $T_\nu^{(1)}$ whenever it exists. As discussed in Section 1, alternative to the traditional limiting distribution calibration approach, it is possible to compute the critical values via the Monte Carlo simulation in an appropriate setting if we are able to acquire at least partial information on the unknown dependence structure from data. Under the null hypothesis $H_0^{(1)} : \boldsymbol{\mu}_1 = \mathbf{0}$, for $\alpha \in (0, 1)$, we motivate from the multivariate central limit theorem with fixed p the following procedure to calculate critical values $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$:

- Let $\tilde{\boldsymbol{\Sigma}}_1$ be an estimate of $\boldsymbol{\Sigma}_1$ from the sample \mathcal{X}_n , and set $\tilde{\mathbf{R}}_1 = \tilde{\mathbf{D}}_1^{-1/2} \tilde{\boldsymbol{\Sigma}}_1 \tilde{\mathbf{D}}_1^{-1/2}$ with $\tilde{\mathbf{D}}_1 = \text{diag}(\tilde{\boldsymbol{\Sigma}}_1)$. Given \mathcal{X}_n , let $\mathbf{W}_{\text{ns}}^{(1)} \sim N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_1)$ and $\mathbf{W}_s^{(1)} \sim N(\mathbf{0}, \tilde{\mathbf{R}}_1)$ be two multivariate Gaussian random vectors, the critical values $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$ can be computed by the conditional $(1 - \alpha)$ -quantiles of $|\mathbf{W}_{\text{ns}}^{(1)}|_\infty$ and $|\mathbf{W}_s^{(1)}|_\infty$; that is,

$$\begin{aligned} \text{cv}_{\text{ns},\alpha}^{(1)} &= \inf \{t \in \mathbb{R} : \mathbb{P}(|\mathbf{W}_{\text{ns}}^{(1)}|_\infty > t \mid \mathcal{X}_n) \leq \alpha\}, \\ \text{cv}_{s,\alpha}^{(1)} &= \inf \{t \in \mathbb{R} : \mathbb{P}(|\mathbf{W}_s^{(1)}|_\infty > t \mid \mathcal{X}_n) \leq \alpha\}. \end{aligned} \quad (2.2)$$

In practice, we can use the Monte Carlo sample quantiles to approximate $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$. More specifically, let $\{\mathbf{W}_{\text{ns},1}, \dots, \mathbf{W}_{\text{ns},M}\}$ and $\{\mathbf{W}_{s,1}, \dots, \mathbf{W}_{s,M}\}$ be random samples independently drawn from $N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_1)$ and $N(\mathbf{0}, \tilde{\mathbf{R}}_1)$, respectively. Then, $\text{cv}_{\text{ns},\alpha}^{(1)}$ and $\text{cv}_{s,\alpha}^{(1)}$ can be estimated by

$$\begin{aligned} \hat{\text{cv}}_{\text{ns},\alpha}^{(1)} &= \inf \{t \in \mathbb{R} : \hat{F}_{\text{ns},M}^{(1)}(t) \geq 1 - \alpha\}, \\ \hat{\text{cv}}_{s,\alpha}^{(1)} &= \inf \{t \in \mathbb{R} : \hat{F}_{s,M}^{(1)}(t) \geq 1 - \alpha\}, \end{aligned}$$

where $\hat{F}_{\text{ns},M}^{(1)}(t) = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{\text{ns},\ell}|_\infty \leq t\}$ and $\hat{F}_{s,M}^{(1)}(t) = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{s,\ell}|_\infty \leq t\}$. For $\nu \in \{\text{ns}, s\}$, the empirical version of test $\Psi_{\nu,\alpha}^{(1)}$ is therefore defined by

$$\hat{\Psi}_{\nu,\alpha}^{(1)}(M) = I\{T_\nu^{(1)} > \hat{\text{cv}}_{\nu,\alpha}^{(1)}\}, \quad (2.3)$$

such that the null hypothesis $H_0^{(1)}$ is rejected whenever $\hat{\Psi}_{\nu,\alpha}^{(1)}(M) = 1$.

In Section 2.2, we discuss the constructions of $\tilde{\boldsymbol{\Sigma}}_1$, and therefore $\tilde{\mathbf{R}}_1$ in details, from which the wide applicability of the test (2.3) will be explored. The proposed testing procedures are fully data driven and easily computed, the rationale of which are based on Gaussian approximation that characterizes the closeness in distribution of a random process to certain Gaussian process.

In Section 3, we will show that this simulation-based testing procedure is valid for p growing with sample size n ; indeed p can be as large as $O\{\exp(n^c)\}$ for some $c > 0$.

2.1.2 Two-sample case

The above testing procedures can be naturally extended to two-sample problem (1.2). Consider independent samples $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ and $\mathcal{Y}_m = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$. Analogously to (2.1), we define the two-sample non-studentized and studentized test statistics by

$$T_{\text{ns}}^{(\text{II})} = \max_{1 \leq k \leq p} \sqrt{\frac{nm}{n+m}} |\bar{X}_k - \bar{Y}_k| \quad \text{and} \quad T_{\text{s}}^{(\text{II})} = \max_{1 \leq k \leq p} \sqrt{\frac{nm}{m\hat{\sigma}_{1k}^2 + n\hat{\sigma}_{2k}^2}} |\bar{X}_k - \bar{Y}_k|, \quad (2.4)$$

respectively, where $\bar{X}_k = n^{-1} \sum_{i=1}^n X_{ik}$, $\bar{Y}_k = m^{-1} \sum_{j=1}^m Y_{jk}$, $\hat{\sigma}_{1k}^2 = n^{-1} \sum_{i=1}^n (X_{ik} - \bar{X}_k)^2$, and $\hat{\sigma}_{2k}^2 = m^{-1} \sum_{j=1}^m (Y_{jk} - \bar{Y}_k)^2$. Intuitively, large values of $T_{\text{ns}}^{(\text{II})}$ or $T_{\text{s}}^{(\text{II})}$ lead to a rejection of the null hypothesis $H_0^{(\text{II})} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. For nominal significance level α , we define α -level tests of the form

$$\Psi_{\text{ns},\alpha}^{(\text{II})} = I\{T_{\text{ns}}^{(\text{II})} > \text{cv}_{\text{ns},\alpha}^{(\text{II})}\} \quad \text{or} \quad \Psi_{\text{s},\alpha}^{(\text{II})} = I\{T_{\text{s}}^{(\text{II})} > \text{cv}_{\text{s},\alpha}^{(\text{II})}\}$$

with appropriate critical values $\text{cv}_{\text{ns},\alpha}^{(\text{II})}$ and $\text{cv}_{\text{s},\alpha}^{(\text{II})}$.

Let $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ be estimates of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ from the samples \mathcal{X}_n and \mathcal{Y}_m , respectively. Define

$$\tilde{\boldsymbol{\Sigma}}_{1,2} = \frac{m}{N} \tilde{\boldsymbol{\Sigma}}_1 + \frac{n}{N} \tilde{\boldsymbol{\Sigma}}_2, \quad \tilde{\mathbf{D}}_{1,2} = \text{diag}(\tilde{\boldsymbol{\Sigma}}_{1,2}), \quad \tilde{\mathbf{R}}_{1,2} = \tilde{\mathbf{D}}_{1,2}^{-1/2} \tilde{\boldsymbol{\Sigma}}_{1,2} \tilde{\mathbf{D}}_{1,2}^{-1/2}, \quad (2.5)$$

and let $\mathbf{W}_{\text{ns}}^{(\text{II})} \sim N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{1,2})$ and $\mathbf{W}_{\text{s}}^{(\text{II})} \sim N(\mathbf{0}, \tilde{\mathbf{R}}_{1,2})$. As in Section 2.1.1, the critical values are taken to be the following conditional $(1 - \alpha)$ -quantiles

$$\begin{aligned} \text{cv}_{\text{ns},\alpha}^{(\text{II})} &= \inf \{t \in \mathbb{R} : \mathbb{P}(|\mathbf{W}_{\text{ns}}^{(\text{II})}|_\infty > t | \mathcal{X}_n, \mathcal{Y}_m) \leq \alpha\}, \\ \text{cv}_{\text{s},\alpha}^{(\text{II})} &= \inf \{t \in \mathbb{R} : \mathbb{P}(|\mathbf{W}_{\text{s}}^{(\text{II})}|_\infty > t | \mathcal{X}_n, \mathcal{Y}_m) \leq \alpha\}, \end{aligned}$$

which can be computed via the Monte Carlo simulations. Let $\{\mathbf{W}_{\text{ns},1}, \dots, \mathbf{W}_{\text{ns},M}\}$ and $\{\mathbf{W}_{\text{s},1}, \dots, \mathbf{W}_{\text{s},M}\}$ be two random samples independently generated from $N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{1,2})$ and $N(\mathbf{0}, \tilde{\mathbf{R}}_{1,2})$, respectively. Then, $\text{cv}_{\text{ns},\alpha}^{(\text{II})}$ and $\text{cv}_{\text{s},\alpha}^{(\text{II})}$ can be estimated by

$$\begin{aligned} \hat{\text{cv}}_{\text{ns},\alpha}^{(\text{II})} &= \inf \{t \in \mathbb{R} : \hat{F}_{\text{ns},M}^{(\text{II})}(t) \geq 1 - \alpha\}, \\ \hat{\text{cv}}_{\text{s},\alpha}^{(\text{II})} &= \inf \{t \in \mathbb{R} : \hat{F}_{\text{s},M}^{(\text{II})}(t) \geq 1 - \alpha\}, \end{aligned}$$

where $\hat{F}_{\text{ns},M}^{(\text{II})}(t) = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{\text{ns},\ell}|_\infty \leq t\}$ and $\hat{F}_{\text{s},M}^{(\text{II})}(t) = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{\text{s},\ell}|_\infty \leq t\}$. Similarly to (2.3), for $\nu \in \{\text{ns}, \text{s}\}$, we define the empirical version of $\Psi_{\nu,\alpha}^{(\text{II})}$ by

$$\hat{\Psi}_{\nu,\alpha}^{(\text{II})}(M) = I\{T_\nu^{(\text{II})} > \hat{\text{cv}}_{\nu,\alpha}^{(\text{II})}\}, \quad (2.6)$$

such that the null hypothesis $H_0^{(\text{II})}$ is rejected as long as $\hat{\Psi}_{\nu,\alpha}^{(\text{II})}(M) = 1$.

2.2 Estimation of covariance matrices

As a part of proposed testing procedures, we need estimates of the covariance matrices. Many existing tests rely on the operator-norm consistent estimation of the covariance matrices that requires extra structural assumptions on the unknown covariances such as banding or sparsity. Under these structural assumptions, regularized or rate-optimal estimates of covariance matrices have been developed (Bickel and Levina, 2008a,b; Cai and Liu, 2011; Cai and Zhou, 2012; Cai and Yuan, 2012). and employed for building tests.

In contrast, the proposed tests require much less restrictions on covariance estimates, which grants its wide scope of applicability. In fact, the validity of the simulation-based testing procedures only entails the covariance estimators $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ to satisfy $\|\tilde{\Sigma}_1 - \Sigma_1\|_\infty = o_P(1)$ and $\|\tilde{\Sigma}_2 - \Sigma_2\|_\infty = o_P(1)$, which is shown in Propositions ?? and ?? included in the supplementary material. Denote by $\hat{\Sigma}_1 = (\hat{\sigma}_{1,k\ell})_{1 \leq k, \ell \leq p}$ and $\hat{\Sigma}_2 = (\hat{\sigma}_{2,k\ell})_{1 \leq k, \ell \leq p}$ the sample covariance matrices based on \mathcal{X}_n and \mathcal{Y}_m , respectively. The corresponding sample correlation matrices are $\hat{\mathbf{R}}_q = \hat{\mathbf{D}}_q^{-1/2} \hat{\Sigma}_q \hat{\mathbf{D}}_q^{-1/2}$, where $\hat{\mathbf{D}}_q = \text{diag}(\hat{\Sigma}_q)$ for $q = 1, 2$. Matrices $\hat{\Sigma}_{1,2}$ and $\hat{\mathbf{R}}_{1,2}$ are defined in the same way that leads to $\tilde{\Sigma}_{1,2}$ and $\tilde{\mathbf{R}}_{1,2}$ as in (2.5), with $\tilde{\Sigma}_1, \tilde{\Sigma}_2$ replaced by $\hat{\Sigma}_1, \hat{\Sigma}_2$. Let $\mathbf{U} = (U_1, \dots, U_p)' = \mathbf{D}_1^{-1/2} \mathbf{X}$, where $\mathbf{D}_1 = \text{diag}(\Sigma_1)$. It will be shown in Lemma ?? in the supplementary material that

- if there exist constants $K_1 > 0$, $K_2 > 1$ and $0 < \gamma \leq 2$ such that

$$\max_{1 \leq k \leq p} \mathbb{E} \exp(K_1 |U_k|^\gamma) \leq K_2 \quad \text{and} \quad \log(p) = o(n^{\gamma/2}),$$

then $\|\hat{\Sigma}_1 - \Sigma_1\|_\infty + \|\hat{\mathbf{R}}_1 - \mathbf{R}_1\|_\infty = o_P(1)$; and

- if there exist constants $r \geq 4$, $\delta > 0$ and $C_1, C_2 > 0$ such that

$$\max_{1 \leq k \leq p} (\mathbb{E} |U_k|^r)^{1/r} \leq C_1 \quad \text{and} \quad p \leq C_2 n^{r/2-1-\delta},$$

then $\|\hat{\Sigma}_1 - \Sigma_1\|_\infty + \|\hat{\mathbf{R}}_1 - \mathbf{R}_1\|_\infty = o_P(1)$.

Therefore, the standard and most natural estimator, the sample covariance matrix can be directly employed by proposed tests.

In particular, for the lower order moment case where the r th moments of all the coordinates of \mathbf{U} are uniformly bounded, the convergence rates of $\|\hat{\Sigma}_1 - \Sigma_1\|_\infty$ and $\|\hat{\mathbf{R}}_1 - \mathbf{R}_1\|_\infty$ are essentially determined by $pn^{1-r/2}$, up to logarithmic factors. Under suitable moment conditions on both \mathbf{X} and \mathbf{Y} , similar results hold for $\hat{\Sigma}_{1,2}$ and $\hat{\mathbf{R}}_{1,2}$. Therefore, the sample covariance and correlation

matrices can be directly used in the proposed testing procedures, while the dimension p is allowed to be as large as either $O\{\exp(n^{c_1})\}$ or $O(n^{c_2})$ for some $c_1, c_2 > 0$, depending on the light or heavy tailedness of the data. In comparison to the existing tests, we do not enforce any structural assumptions on the covariance matrices Σ_1 and Σ_2 for the simulation-based tests. This echoes our motivations in Section 1 and grants the proposed methods a much wider scope of applicability in practice. As evidenced by extensive numerical studies in Section 4, our proposed procedures are fairly robust to various covariance structures with complex forms, even the long range dependence.

Although the proposed tests do not require operator-norm consistent estimates of Σ_1 and Σ_2 , still one may replace the sample covariance matrix by adaptive and rate-optimal covariance estimators to improve the empirical performance when the underlying covariance satisfies certain structural assumptions. In practice, an additional eigenvalue correction step might be employed to enforce the positive semi-definiteness of the covariance estimator as discussed in Cai and Liu (2011).

2.3 Screening-based testing procedures

The proposed testing procedures are valid when the dimension p is much larger than the sample size n . However, building tests based on all dimensions may result in large cut-off values which will potentially compromise the power performance. To enhance the power, we propose a two-step procedure that combines the proposed simulation-based tests and a preliminary step on *feature screening*, which screens the p measurements before conducting the test. The power of this two-step procedure is expected to improve upon the proposed tests with a large number of irrelevant features excluded. Numerical experiments in Section 4 provide evidences that the two-step testing procedures substantially improve the power performance.

We take the non-studentized statistic $T_{\text{ns}}^{(1)}$ in the one-sample problem to illustrate the motivation of the two-step procedure. Under sparse alternatives, assume that irrelevant coordinates with $\mu_{1k} = 0$ have been identified and removed from constructing testing statistic. Denote by $\widehat{\mathcal{S}}$ the index set of excluded coordinates upon screening. The analogue of $T_{\text{ns}}^{(1)}$ is given by $T_{\text{ns}}^{f,(1)} = \max_{k \notin \widehat{\mathcal{S}}} \sqrt{n} |\bar{X}_k|$. As $\{1, \dots, p\} \setminus \widehat{\mathcal{S}} \subset \{1, \dots, p\}$, the critical value for $T_{\text{ns}}^{f,(1)}$ at a given significance level α is no larger than $\text{cv}_{\text{ns},\alpha}^{(1)}$, the critical value for the original test statistic $T_{\text{ns}}^{(1)}$. For suitably selected $\widehat{\mathcal{S}}$, $T_{\text{ns}}^{f,(1)}$ coincides with $T_{\text{ns}}^{(1)}$ with high probability. Therefore, the two-step procedure will enhance the power in comparison to the original test. As shown in Section 3, two-step procedures maintain the nominal significance level and are consistent against sparse alternatives.

2.3.1 One-sample case

Given $\mu_1 = (\mu_{11}, \dots, \mu_{1p})'$, let $\mathcal{S}_{10} = \{k = 1, \dots, p : \mu_{1k} = 0\}$. The preliminary procedure is aimed at eliminating irrelevant features indexed by \mathcal{S}_{10} . Reformulate the original global test of a

mean vector to the following p marginal tests:

$$H_{0k}^{(I)} : \mu_{1k} = 0 \text{ versus } H_{1k}^{(I)} : \mu_{1k} \neq 0,$$

for $k = 1, \dots, p$. For the k th marginal hypothesis, a standard test statistic is the t -statistic $\text{TS}_k^{(I)} = \sqrt{n}\bar{X}_k/\hat{\sigma}_{1k}$. Motivated by the idea of marginal screening (Chang, Tang and Wu, 2013), we define the following index set

$$\widehat{\mathcal{S}}_1 = \left\{ 1 \leq k \leq p : |\text{TS}_k^{(I)}| \leq \left[\sqrt{2} + \frac{\sqrt{2}}{2\log(p)} + \sqrt{\frac{2\log(1/\alpha)}{\log(p)}} \right] \sqrt{\log(p)} \right\}. \quad (2.7)$$

We refer to Chang, Tang and Wu (2013) for more discussions on the advantages of the studentized statistics in marginal screening problems. If $|\widehat{\mathcal{S}}_1| < p$, we put $d = p - |\widehat{\mathcal{S}}_1|$ and let $\tilde{\boldsymbol{\mu}}_1 \in \mathbb{R}^d$ be the sub-vector of $\boldsymbol{\mu}_1 \in \mathbb{R}^p$ containing only the coordinates excluded by $\widehat{\mathcal{S}}_1$. We have therefore downsized the original problem and instead, we focus on the reduced null hypothesis $\tilde{H}_0^{(I)} : \tilde{\boldsymbol{\mu}}_1 = \mathbf{0}$ against the alternative $\tilde{H}_1^{(I)} : \tilde{\boldsymbol{\mu}}_1 \neq \mathbf{0}$. The resulting non-studentized and studentized tests are given by

$$\Psi_{\text{ns},\alpha}^{f,(I)} = I \left\{ \max_{k \notin \widehat{\mathcal{S}}_1} \sqrt{n} |\bar{X}_k| > \text{cv}_{\text{ns},\alpha}^{(I)}(\widehat{\mathcal{S}}_1) \right\} \quad \text{and} \quad \Psi_{\text{s},\alpha}^{f,(I)} = I \left\{ \max_{k \notin \widehat{\mathcal{S}}_1} \frac{\sqrt{n} |\bar{X}_k|}{\hat{\sigma}_{1k}} > \text{cv}_{\text{s},\alpha}^{(I)}(\widehat{\mathcal{S}}_1) \right\}, \quad (2.8)$$

where $\text{cv}_{\text{ns},\alpha}^{(I)}(\widehat{\mathcal{S}}_1)$ and $\text{cv}_{\text{s},\alpha}^{(I)}(\widehat{\mathcal{S}}_1)$ denote the conditional $(1 - \alpha)$ -quantile of $\max_{k \notin \widehat{\mathcal{S}}_1} |W_{\text{ns},k}^{(I)}|$ and $\max_{k \notin \widehat{\mathcal{S}}_1} |W_{\text{s},k}^{(I)}|$ given \mathcal{X}_n , respectively, where $\mathbf{W}_{\text{ns}}^{(I)} = (W_{\text{ns},1}^{(I)}, \dots, W_{\text{ns},p}^{(I)})'$ and $\mathbf{W}_{\text{s}}^{(I)} = (W_{\text{s},1}^{(I)}, \dots, W_{\text{s},p}^{(I)})'$ as in (2.2). If $|\widehat{\mathcal{S}}_1| = p$, we set $\Psi_{\text{ns},\alpha}^{f,(I)} = \Psi_{\text{s},\alpha}^{f,(I)} = 0$.

2.3.2 Two-sample case

Similar to the one-sample case, for each $k = 1, \dots, p$, we define $\text{TS}_k^{(\text{II})} = \sqrt{nm} |\bar{X}_k - \bar{Y}_k| / (m\hat{\sigma}_{1k}^2 + n\hat{\sigma}_{2k}^2)^{1/2}$ and set

$$\widehat{\mathcal{S}}_2 = \left\{ 1 \leq k \leq p : |\text{TS}_k^{(\text{II})}| \leq \left[\sqrt{2} + \frac{\sqrt{2}}{2\log(p)} + \sqrt{\frac{2\log(1/\alpha)}{\log(p)}} \right] \sqrt{\log(p)} \right\}. \quad (2.9)$$

If $|\widehat{\mathcal{S}}_2| < p$, we put $d = p - |\widehat{\mathcal{S}}_2|$ and for $q = 1, 2$, let $\tilde{\boldsymbol{\mu}}_q \in \mathbb{R}^d$ be the sub-vector of $\boldsymbol{\mu}_q \in \mathbb{R}^p$ indexed by $\{1, \dots, p\} \setminus \widehat{\mathcal{S}}_2$. Then we carry out the testing procedures in Section 2.1.2 to the reduced problem: $\tilde{H}_0^{(\text{II})} : \tilde{\boldsymbol{\mu}}_1 = \tilde{\boldsymbol{\mu}}_2$ against the alternative $\tilde{H}_1^{(\text{II})} : \tilde{\boldsymbol{\mu}}_1 \neq \tilde{\boldsymbol{\mu}}_2$. The resulting tests, denoted by $\Psi_{\text{ns},\alpha}^{f,(\text{II})}$ and $\Psi_{\text{s},\alpha}^{f,(\text{II})}$, are therefore defined in the same way as $\Psi_{\text{ns},\alpha}^{f,(I)}$ and $\Psi_{\text{s},\alpha}^{f,(I)}$ in (2.8) respectively. If $|\widehat{\mathcal{S}}_2| = p$, we set $\Psi_{\text{ns},\alpha}^{f,(\text{II})} = \Psi_{\text{s},\alpha}^{f,(\text{II})} = 0$.

3 Theoretical properties

In this section, we study the properties of the proposed tests including the asymptotic sizes and powers. In practice, taking M in thousands using numerical devices to increase simulation efficiency is now the rule rather than the exception in the Monte Carlo framework. The difference between such large values of M and using mathematically ideal value $M = \infty$ is particularly small. We therefore focus on the oracle tests $\Psi_{\nu,\alpha}^{(I)}$ and $\Psi_{\nu,\alpha}^{(II)}$ for $\nu \in \{\text{ns}, \text{s}\}$, and their screening-based analogues $\Psi_{\nu,\alpha}^{f,(I)}$ and $\Psi_{\nu,\alpha}^{f,(II)}$. It is shown that the proposed tests maintain the nominal size asymptotically under very general covariance structures. Moreover, the proposed tests are shown to be consistent against sparse alternatives.

For p -dimensional random vectors $\mathbf{X} = (X_1, \dots, X_p)'$ and $\mathbf{Y} = (Y_1, \dots, Y_p)'$ with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, their marginally standardized versions are defined by $\mathbf{U} = (U_1, \dots, U_p)' = \mathbf{D}_1^{-1/2}\mathbf{X}$ and $\mathbf{V} = (V_1, \dots, V_p)' = \mathbf{D}_2^{-1/2}\mathbf{Y}$, respectively, where $\mathbf{D}_1 = \text{diag}(\boldsymbol{\Sigma}_1)$ and $\mathbf{D}_2 = \text{diag}(\boldsymbol{\Sigma}_2)$. As discussed in Section 2, no structural assumptions on the unknown covariance matrices are enforced. We only impose the following mild moment conditions.

(M1) (Lower order moments) There exist $r \geq 4$ and $K_0 > 0$ such that the r th moments of the components of \mathbf{U} and \mathbf{V} are uniformly bounded; that is,

$$\max_{1 \leq k \leq p} (\mathbb{E}|U_k|^r)^{1/r} \leq K_0 \quad \text{and} \quad \max_{1 \leq k \leq p} (\mathbb{E}|V_k|^r)^{1/r} \leq K_0.$$

(M2) (Sub-exponential tails) There exist constants $K_1 > 0$, $K_2 > 1$ and $0 < \gamma \leq 2$ such that

$$\max_{1 \leq k \leq p} \mathbb{E} \exp(K_1 |U_k|^\gamma) \leq K_2 \quad \text{and} \quad \max_{1 \leq k \leq p} \mathbb{E} \exp(K_1 |V_k|^\gamma) \leq K_2.$$

Throughout this section, we assume that $n, p \geq 2$, $n \asymp m$, $n \leq m$, and that $\{\sigma_{qk}^2 := \sigma_{q,kk}, q = 1, 2\}_{k=1}^n$ is bounded away from 0 and ∞ .

Theorem 1 (Asymptotic size of the one-sample tests without screening). *Let $\tilde{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_1$, the sample covariance matrix, in (2.2) and $\nu \in \{\text{ns}, \text{s}\}$.*

(i) *Assume that (M1) holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$. Then as $n, p \rightarrow \infty$,*

$$\mathbb{P}_{H_0^{(I)}} \{ \Psi_{\nu,\alpha}^{(I)} = 1 \} \rightarrow \alpha.$$

(ii) *Assume that (M2) holds for some $\gamma \geq \frac{1}{2}$ and $\log(p) = o(n^{1/7})$. Then as $n, p \rightarrow \infty$,*

$$\mathbb{P}_{H_0^{(I)}} \{ \Psi_{\nu,\alpha}^{(I)} = 1 \} \rightarrow \alpha.$$

Theorem 1 establishes the validity of the proposed one-sample tests in the sense that the testing procedures in Section 2.1.1 maintain nominal significance level asymptotically. In addition, as evidenced by the numerical experiments in Section 4, the test based on non-studentized statistics outperforms its studentized analogue in terms of maintaining the nominal significance level when the sample size is small. This, however, is not surprising since the inverse operation, say $\widehat{\mathbf{D}}_1^{-1/2}$, usually leads to an augmentation of the estimation error in $\widehat{\mathbf{D}}_1$ and therefore is more sensitive to the sample size. In the following theorem, we summarize the asymptotic power of the proposed one-sample tests under suitable conditions on the lower bound of the signal-to-noise ratios.

Theorem 2 (Asymptotic power of the one-sample tests without screening). *Let $\tilde{\Sigma}_1 = \widehat{\Sigma}_1$, the sample covariance matrix, in (2.2). Assume that either condition (M1) holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition (M2) holds and $\log(p) = o(n^{\gamma/2})$. For given $0 < \alpha < 1$, write*

$$\lambda(p, \alpha) = \sqrt{2 \log(p)} + \sqrt{2 \log(1/\alpha)}, \quad (3.1)$$

and let $\{\varepsilon_n\}_{n \geq 1}$ be an arbitrary sequence of positive numbers satisfying $\varepsilon_n \rightarrow 0$ and $\varepsilon_n \sqrt{\log(p)} \rightarrow \infty$ as $n \rightarrow \infty$.

(i) Under the alternative $H_1^{(I)}$ with

$$\frac{\max_{1 \leq k \leq p} |\mu_{1k}|}{\max_{1 \leq k \leq p} \sigma_{1k}} \geq (1 + \varepsilon_n) n^{-1/2} \lambda(p, \alpha),$$

we have as $n, p \rightarrow \infty$,

$$\mathbb{P}_{H_1^{(I)}} \{ \Psi_{\text{ns}, \alpha}^{(I)} = 1 \} \rightarrow 1.$$

(ii) Under the alternative $H_1^{(I)}$ with

$$\max_{1 \leq k \leq p} \frac{|\mu_{1k}|}{\sigma_{1k}} \geq (1 + \varepsilon_n) n^{-1/2} \lambda(p, \alpha),$$

we have as $n, p \rightarrow \infty$,

$$\mathbb{P}_{H_1^{(I)}} \{ \Psi_{\text{s}, \alpha}^{(I)} = 1 \} \rightarrow 1.$$

Theorem 2 shows that, in agreement with intuition, the test based on studentized statistics is consistent in a larger testable region in comparison to the test based on non-studentized statistics. As a complement to Theorem 1, the asymptotic size of the proposed two-sample tests without screening is reported below.

Theorem 3 (Asymptotic size of the two-sample tests without screening). *Let $(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = (\widehat{\Sigma}_1, \widehat{\Sigma}_2)$, which are the sample covariance matrices, and $\nu \in \{\text{ns}, \text{s}\}$. Assume that either condition (i) or condition (ii) in Theorem 1 holds. Then as $n, p \rightarrow \infty$,*

$$\mathbb{P}_{H_0^{(\text{II})}} \{ \Psi_{\nu, \alpha}^{(\text{II})} = 1 \} \rightarrow \alpha.$$

Theorem 3 implies that, under proper moment conditions, the proposed two-sample non-screening tests maintain nominal size α asymptotically, while allowing for either a polynomial or an exponential rate of growth of the dimension p with respect to the sample size n . In Theorem 4 below, the asymptotic power of the two-sample non-screening tests is analyzed under conditions on the separation distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Theorem 4 (Asymptotic power of the two-sample tests without screening). *Let $(\tilde{\boldsymbol{\Sigma}}_1, \tilde{\boldsymbol{\Sigma}}_2) = (\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$, which are the sample covariance matrices. Assume that either condition (M1) holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition (M2) holds and $\log(p) = o(n^{\gamma/2})$. For given $0 < \alpha < 1$, let $\lambda(p, \alpha)$ be as in (3.1) and let $\{\varepsilon_n\}_{n \geq 1}$ be an arbitrary sequence of positive numbers satisfying $\varepsilon_n \rightarrow 0$ and $\varepsilon_n \sqrt{\log(p)} \rightarrow \infty$ as $n \rightarrow \infty$.*

(i) Under the alternative $H_1^{(\text{II})}$ with

$$\frac{\max_{1 \leq k \leq p} |\mu_{1k} - \mu_{2k}|}{\max_{1 \leq k \leq p} (\sigma_{1k}^2/n + \sigma_{2k}^2/m)^{1/2}} \geq (1 + \varepsilon_n) \lambda(p, \alpha),$$

we have as $n, p \rightarrow \infty$,

$$\mathbb{P}_{H_1^{(\text{II})}} \{ \Psi_{\text{ns}, \alpha}^{(\text{II})} = 1 \} \rightarrow 1.$$

(ii) Under the alternative $H_1^{(\text{II})}$ with

$$\max_{1 \leq k \leq p} \frac{|\mu_{1k} - \mu_{2k}|}{(\sigma_{1k}^2/n + \sigma_{2k}^2/m)^{1/2}} \geq (1 + \varepsilon_n) \lambda(p, \alpha),$$

we have as $n, p \rightarrow \infty$,

$$\mathbb{P}_{H_1^{(\text{II})}} \{ \Psi_{\text{s}, \alpha}^{(\text{II})} = 1 \} \rightarrow 1.$$

The following theorem establishes asymptotic properties of the proposed two-step testing procedure. Part (i) in Theorem 5 below shows that the type I error of the proposed screening-based two-step procedures, with or without studentization, can be controlled by the prescribed significance level asymptotically. Similar to the comparison between the studentized and non-studentized tests in Theorem 2, parts (ii) and (iii) in Theorem 5 below also imply that the screening-based two-step studentized test is consistent in a larger region than its non-studentized counterpart.

Theorem 5 (Asymptotic properties of the one-sample screening-based tests). *Let $\tilde{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_1$, the sample covariance matrix, in (2.2). Assume that either condition (M1) holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition (M2) holds for some $\gamma \geq \frac{1}{2}$ and $\log(p) = o(n^{1/7})$.*

(i) Under the null $H_0^{(\text{I})}$, we have $\limsup_{n \rightarrow \infty} \mathbb{P}_{H_0^{(\text{I})}} \{ \Psi_{\nu, \alpha}^{f, (\text{I})} = 1 \} \leq \alpha$ for $\nu \in \{\text{ns}, \text{s}\}$.

(ii) Under the alternative $H_1^{(\text{I})}$ specified in part (i) of Theorem 2, we have $\mathbb{P}_{H_1^{(\text{I})}} \{ \Psi_{\text{ns}, \alpha}^{f, (\text{I})} = 1 \} \rightarrow 1$.

(iii) Under the alternative $H_1^{(I)}$ specified in part (ii) of Theorem 2, we have $\mathbb{P}_{H_1^{(I)}}\{\Psi_{s,\alpha}^{f,(I)} = 1\} \rightarrow 1$.

Analogously, the following theorem establishes the limiting null property and the asymptotic power for the proposed two-step procedures with pre-screening in the two-sample settings. The proof of Theorem 6 is almost identical to that of Theorem 5, and therefore is omitted in supplementary material.

Theorem 6 (Asymptotic properties of the two-sample screening-based tests). *Let $(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = (\hat{\Sigma}_1, \hat{\Sigma}_2)$, which are the sample covariance matrices. Assume that either condition (M1) holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition (M2) holds for some $\gamma \geq \frac{1}{2}$ and $\log(p) = o(n^{1/7})$.*

- (i) Under the null $H_0^{(II)}$, we have $\limsup_{n \rightarrow \infty} \mathbb{P}_{H_0^{(II)}}\{\Psi_{\nu,\alpha}^{f,(II)} = 1\} \leq \alpha$ for $\nu \in \{\text{ns}, s\}$.
- (ii) Under the alternative $H_1^{(II)}$ specified in part (i) of Theorem 4, we have $\mathbb{P}_{H_1^{(II)}}\{\Psi_{\text{ns},\alpha}^{f,(II)} = 1\} \rightarrow 1$.
- (iii) Under the alternative $H_1^{(II)}$ specified in part (ii) of Theorem 4, we have $\mathbb{P}_{H_1^{(II)}}\{\Psi_{s,\alpha}^{f,(II)} = 1\} \rightarrow 1$.

4 Simulation studies

In this section, we report the simulation results from several experiments which were designed to evaluate the performance of the proposed tests, including the non-studentized test without screening $\Psi_{\text{ns},\alpha}$, the studentized test without screening $\Psi_{s,\alpha}$, the non-studentized test with screening $\Psi_{\text{ns},\alpha}^f$ and the studentized test with screening $\Psi_{s,\alpha}^f$, for both one- and two-sample problems. For ease of exposition, we suppress the superscripts (I) and (II). To demonstrate the proposed tests, we also implemented peer testing procedures for comparison. For the one-sample problem, we compared the proposed tests with the test by Zhong, Chen and Xu (2013) (denoted by ZCX hereafter) and the Higher Criticism procedure by Donoho and Jin (2004) (denoted by HC hereafter). For the two-sample problem, we experimented three existing tests: the one by Chen and Qin (2010) (denoted by CQ hereafter), the test based on the Higher Criticism procedure Delaigle, Hall and Jin (2011) (denoted by HC2 hereafter), and the test by Cai, Liu and Xia (2014) (denoted by CLX hereafter).

The proposed tests are easily implementable via the Monte Carlo method. In the simulation studies, we considered a wide range of covariance structures, including both the sparse and dense settings to investigate the numerical performance of the proposed tests, as well as different values of sample sizes (n, m) and dimension p . As discussed in Sections 2.2 and 3, the validity of the proposed tests is guaranteed as long as the covariance estimators are uniform entrywise consistent. Therefore, we used the sample covariance matrices to generate $M = 1500$ Monte Carlo samples to compute the empirical critical values for the tests. The simulation results show that the proposed

tests reasonably maintain the nominal significance level, and have substantially better power performance in comparison to peer tests for weak and/or sparse signals.

4.1 One-sample case

Without loss of generality, we took $\boldsymbol{\mu}_1 = \mathbf{0}$ under the null hypothesis for the one-sample problem (1.1), whereas, under the alternative, we took $\boldsymbol{\mu}_1 = (\mu_{11}, \dots, \mu_{1p})'$ to have $\lfloor \kappa p^r \rfloor$ non-zero entries uniformly and randomly drawn from $\{1, \dots, p\}$, where κ was an integer and $\lfloor x \rfloor$ denotes the integer part of x . We took $r = 0, 0.4, 0.5, 0.7$ and 0.85 , where $\kappa = 8$ if $r = 0$ and $\kappa = 1$ otherwise. The settings $r = 0$ and $r = 0.7$ or $r = 0.85$ mimic the scenarios of sparse and non-sparse signals, respectively. The magnitudes of non-zero entries $\mu_{1\ell}$ were set to be $\{2\beta\sigma_{1,\ell\ell} \log(p)/n\}^{1/2}$, where $\sigma_{1,\ell\ell}$ denotes the ℓ th diagonal entry of $\boldsymbol{\Sigma}_1$. We took $\beta = 0.01, 0.2, 0.4, 0.6$ and $\beta = 0.01$ to mimic the scenario of weak signals.

The data generation models are described as follows. Let $\text{Unif}(a, b)$ be the uniform distribution on (a, b) for $b > a$. The following three models were used to generate random samples $\mathbf{X}_i = \mathbf{W}_i + \boldsymbol{\mu}_1$ for $i = 1, \dots, n$, where $\{\mathbf{W}_i = (W_{i,1}, \dots, W_{i,p})' : i = 1, \dots, n\}$ are independent and identically distributed centered Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1 \leq k, \ell \leq p}$.

- (a) Model 1^(I) (Bandable $\boldsymbol{\Sigma}_1$): $\sigma_{1,k\ell} = 0.4^{|k-\ell|}$ for $1 \leq k, \ell \leq p$. This model was employed by [Zhong, Chen and Xu \(2013\)](#).
- (b) Model 2^(I) (Block diagonal $\boldsymbol{\Sigma}_1$): $\sigma_{1,kk}$ are independent and identically drawn from $\text{Unif}(1, 2)$, $\sigma_{1,k\ell} = 0.7$ for $10(t-1) + 1 \leq k \neq \ell \leq 10t$, where $t = 1, \dots, \lfloor p/10 \rfloor$, and $\sigma_{1,k\ell} = 0$ otherwise.
- (c) Model 3^(I) (Long range dependence $\boldsymbol{\Sigma}_1$): Let $\theta_1, \dots, \theta_p$ be independent and identically drawn from $\text{Unif}(1, 2)$, and we took $\sigma_{1,kk} = \theta_k$ and $\sigma_{1,k\ell} = \rho_\alpha(|k - \ell|)$ for $k \neq \ell$, where $\rho_\alpha(e) = \frac{1}{2}\{(e+1)^{2H} + (e-1)^{2H} - 2e^{2H}\}$ with $H = 0.9$.

Model 1^(I) and Model 2^(I) have sparse covariance structures while Model 3^(I) takes long range dependence into account which possesses a non-sparse structure. In addition, we considered the following two models with non-Gaussian data generation mechanisms to study the robustness of the proposed tests against Gaussian assumptions.

- (d) Model 4^(I) (Autoregressive process of order one, AR(1), with t -distributed innovations): Generate independent and identically distributed p -variate random vectors $\mathbf{X}_i \sim t_\omega(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ for $i = 1, \dots, n$, where $t_\omega(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is the non-central multivariate t -distribution with non-central parameter $\boldsymbol{\mu}_1$, $\omega = 5$ degrees of freedom, and $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1 \leq k, \ell \leq p}$ with $\sigma_{1,k\ell} = 0.995^{|k-\ell|}$.
- (e) Model 5^(I) (Moving average process with Beta distributed innovations): For $i = 1, \dots, n$ and $k = 1, \dots, p$, we considered $X_{ik} = \rho_1 Z_{i,k} + \rho_2 Z_{i,k+1} + \dots + \rho_p Z_{i,k+p-1} + \mu_k$ where ρ_ℓ are

independent and identically drawn from $0.6\text{Unif}(-1, 1) + 0.4\delta_0$ for $\ell = 1, \dots, p$, where δ_0 is the point mass at 0 and $\{Z_{i,k}\}$ are independent random variables with a common centered Beta(2, 1) distribution.

Both the covariance structures in Model 4^(I) and Model 5^(I) are non-sparse. Simulation results show that the proposed tests are valid in these non-sparse models with non-Gaussian sampling distributions.

In each model, we generated data with a sample size $n = 40$ or $n = 80$. The dimension p took values 120, 360 and 1080. The empirical power and level of significance were computed based on 1500 simulations. Simulation results for the tests $\Psi_{\text{ns},\alpha}$, $\Psi_{\text{s},\alpha}$, $\Psi_{\text{ns},\alpha}^f$ and $\Psi_{\text{s},\alpha}^f$ and the ZCX and HC tests are summarized in Table 1 and Figures 1-5. Table 1 displays the empirical sizes of all the tests. It can be seen that in all the models, the empirical sizes of the non-studentized tests $\Psi_{\text{ns},\alpha}$ and $\Psi_{\text{ns},\alpha}^f$ are reasonably close to the nominal level 0.05 for both $n = 40$ and $n = 80$. Comparing the results for $n = 40$ and $n = 80$, the proposed studentized tests $\Psi_{\text{s},\alpha}$ and $\Psi_{\text{s},\alpha}^f$ have slightly inflated size when n is relatively small but improve with larger sample sizes. The ZCX test maintains the nominal size for Model 1^(I) but fails in the presence of long range dependence or non-sparse covariance structures. The HC procedure also fails in maintaining the nominal significance when the sample size n is small or the dependency is strong and complex.

To compare the empirical powers, we took $n = 80$, $p = 1080$. For Model 1^(I), we compared the proposed tests with the ZCX test (Figure 1), whereas, for the other four models, we focused on comparing the proposed tests as they maintain the nominal size reasonably well (Figures 2-5). Figure 1 shows that the proposed studentized tests and non-studentized test with screening, $\Psi_{\text{s},\alpha}$, $\Psi_{\text{s},\alpha}^f$ and $\Psi_{\text{ns},\alpha}^f$, provide non-trivial powers against alternatives with sparse signals ($r = 0$) even under the weak signal settings ($\beta = 0.01$); in contrast, the ZCX test improves its power as the signal getting dense, which is expected for sum of squares-type statistics. As the signal strength increases, all tests under consideration gain powers. The proposed tests with screening, $\Psi_{\text{ns},\alpha}^f$ and $\Psi_{\text{s},\alpha}^f$, outperform the ZCX test under sparse alternatives ($r = 0, 0.4$), and their powers are close to that of the ZCX test for dense signals ($r \geq 0.7$). From Figures 2 to 5, we observe that the screening procedure substantially improves the power performance of the tests for all settings, which reflects the heuristic discussions and motivations in Section 2.3. The proposed non-studentized test with screening $\Psi_{\text{ns},\alpha}^f$ performs comparably to, or better than, the studentized test without screening $\Psi_{\text{s},\alpha}$ under sparse alternatives ($r \leq 0.5$). This suggests that $\Psi_{\text{ns},\alpha}^f$ is more preferable in practice given its capability in maintaining the nominal significance for small sample size. More extensive simulations were carried out for dimensions $p = 120$ and 360, from which the comparisons are in line with the cases reported here. It can be further observed that the empirical powers of all the tests increase in p . These numerical results support our theoretical analysis. The additional simulation results are placed in the supplementary material.

In summary, the numerical results show that the proposed tests, particularly the studentized

		Model 1 ^(t)				Model 2 ^(t)				Model 3 ^(t)				Model 4 ^(t)				Model 5 ^(t)			
tests / p		120	360	1080	120	360	1080	120	360	1080	120	360	1080	120	360	1080	120	360	1080	360	1080
$n = 40$																					
$\Psi_{\text{ns},\alpha}$		0.037	0.027	0.021	0.035	0.033	0.020	0.025	0.028	0.023	0.054	0.044	0.033	0.044	0.023	0.023	0.044	0.023	0.023	0.023	0.023
$\Psi_{\text{s},\alpha}$		0.133	0.126	0.168	0.094	0.114	0.216	0.093	0.113	0.202	0.065	0.080	0.096	0.094	0.190	0.252	0.094	0.190	0.190	0.190	0.252
$\Psi_{\text{ns},\alpha}^f$		0.044	0.045	0.043	0.041	0.045	0.030	0.039	0.027	0.039	0.054	0.046	0.033	0.055	0.030	0.032	0.055	0.030	0.030	0.030	0.032
$\Psi_{\text{s},\alpha}^f$		0.150	0.154	0.194	0.093	0.152	0.226	0.095	0.170	0.218	0.060	0.058	0.093	0.095	0.205	0.305	0.095	0.205	0.205	0.205	0.305
ZCX		0.064	0.078	0.089	1	1	1	1	1	1	0.382	0.487	0.673	0.389	1	1	0.389	1	1	1	1
HC		0.165	0.296	0.545	0.195	0.284	0.510	0.145	0.278	0.418	0.370	0.450	0.535	0.179	0.280	0.518	0.179	0.280	0.280	0.280	0.518
$n = 80$																					
$\Psi_{\text{ns},\alpha}$		0.037	0.036	0.029	0.043	0.032	0.025	0.040	0.032	0.042	0.049	0.047	0.040	0.034	0.039	0.036	0.034	0.039	0.039	0.039	0.036
$\Psi_{\text{s},\alpha}$		0.060	0.082	0.092	0.073	0.092	0.091	0.082	0.083	0.094	0.058	0.058	0.067	0.083	0.090	0.125	0.083	0.090	0.090	0.090	0.125
$\Psi_{\text{ns},\alpha}^f$		0.048	0.045	0.043	0.053	0.043	0.034	0.051	0.045	0.040	0.049	0.048	0.044	0.046	0.054	0.045	0.044	0.054	0.054	0.054	0.045
$\Psi_{\text{s},\alpha}^f$		0.086	0.097	0.094	0.088	0.091	0.106	0.095	0.091	0.110	0.060	0.058	0.069	0.091	0.098	0.157	0.069	0.091	0.098	0.098	0.157
ZCX		0.080	0.072	0.071	1	1	1	1	1	1	0.404	0.506	0.702	0.593	1	1	0.593	1	1	1	1
HC		0.129	0.170	0.247	0.132	0.148	0.250	0.126	0.175	0.223	0.380	0.420	0.506	0.124	0.138	0.283	0.506	0.124	0.138	0.138	0.283

Table 1: Empirical sizes of the proposed tests (non-studentized without screening $\Psi_{\text{ns},\alpha}$, studentized without screening $\Psi_{\text{s},\alpha}$, non-studentized with screening $\Psi_{\text{ns},\alpha}^f$, and studentized with screening $\Psi_{\text{s},\alpha}^f$) for the one-sample problem (1.1), along with those of the tests by [Zhong, Chen and Xu \(2013\)](#) (ZCX) and by [Donoho and Jin \(2004\)](#) (HC) at 5% nominal significance. Models with Gaussian data and bandable, block diagonal or long range dependence covariance matrices, the autoregressive model with t -distributed innovations, and the moving average model with Beta distributed innovations are considered when $n = 40, 80$ and $p = 120, 360, 1080$.

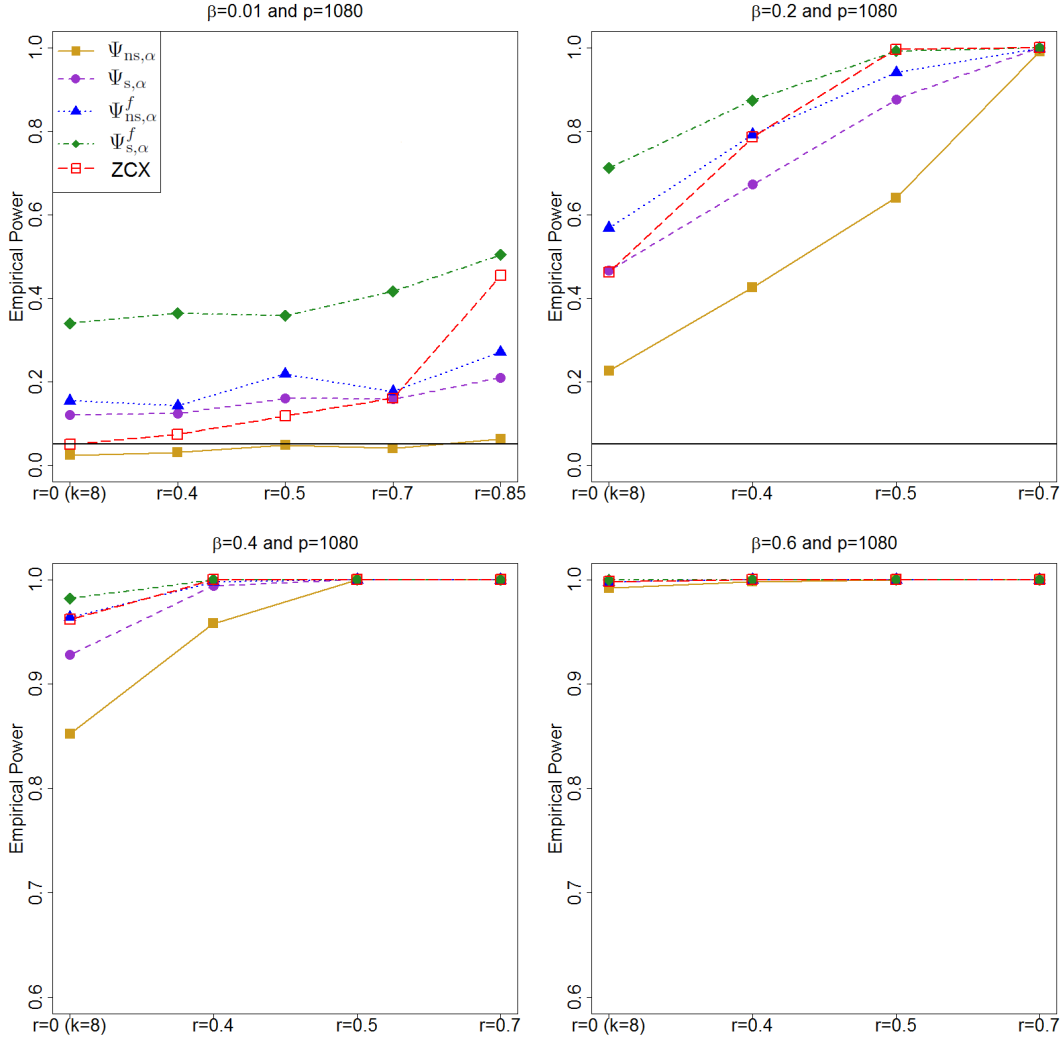


Figure 1: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of the signal strength (β) and sparsity ($1-r$) for the one-sample problem (1.1), along with the power of the test by Zhong, Chen and Xu (2013) (ZCX) at 5% nominal significance for the Gaussian data and bandable covariance matrices in Model 1^(I) when $n = 80$ and $p = 1080$.

tests and the non-studentized test with screening, $\Psi_{s,\alpha}$, $\Psi_{s,\alpha}^f$ and $\Psi_{ns,\alpha}^f$, outperform the peer tests under sparse alternatives when the covariance structure is non-sparse, in other words, when the dependency is spread. The proposed tests are robust against unknown covariance structures and Gaussianity. The non-studentized test with screening $\Psi_{ns,\alpha}^f$ maintains the nominal significance for small sample size and has good powers against sparse alternatives, which is recommended for practical applications with relatively small sample size. The studentized tests with screening $\Psi_{s,\alpha}^f$ is more powerful and thus is preferable in applications with relatively large samples, such as

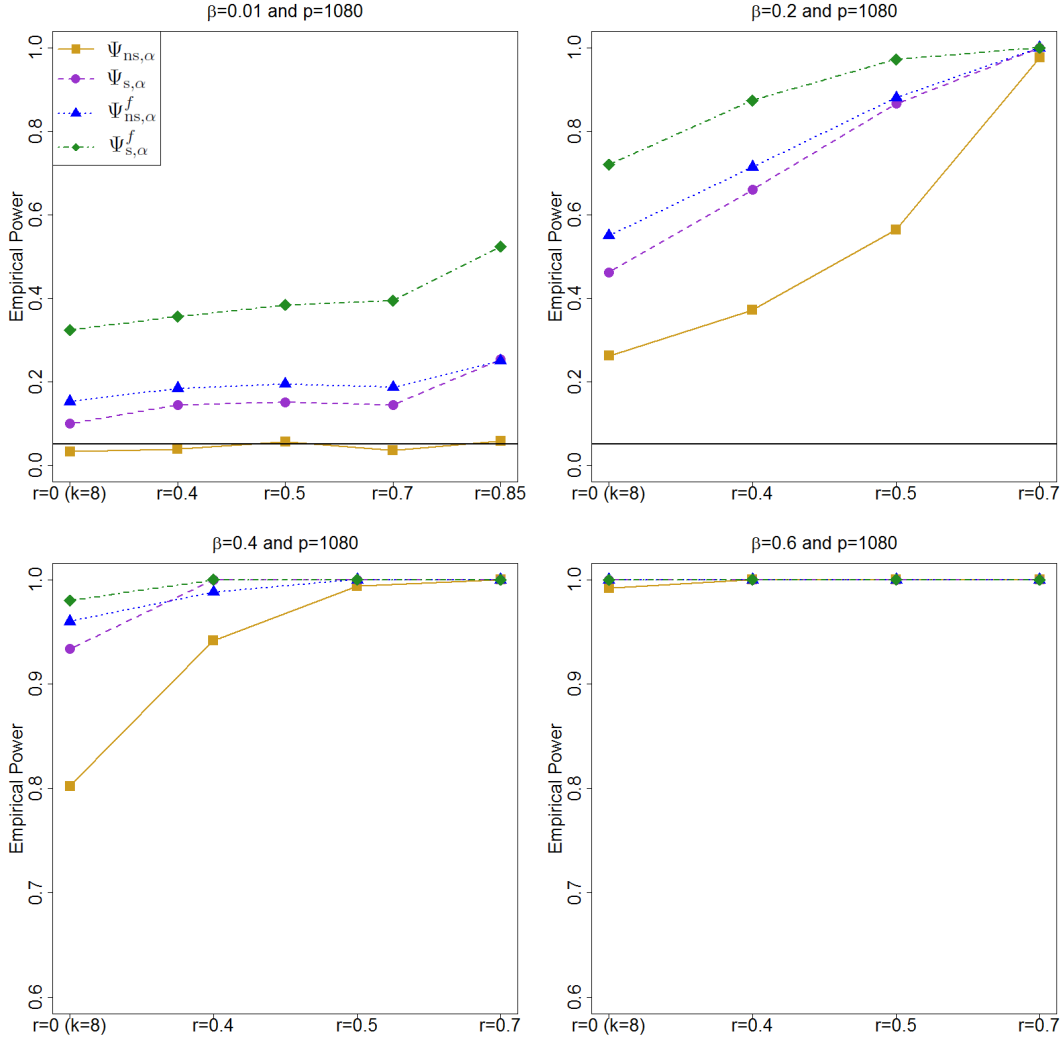


Figure 2: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of the signal strength (β) and sparsity ($1 - r$) for the one-sample problem (1.1) at 5% nominal significance for the Gaussian data and block diagonal covariance matrices in Model 2^(I) when $n = 80$ and $p = 1080$.

biomedical research with a large cohort.

4.2 Two-sample case

Similar settings to those in Section 4.1 were used to examine the proposed tests for the two-sample problem (1.2). Without loss of generality, we took $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ under the null hypothesis, whereas, under the alternative, we let $\boldsymbol{\mu}_1 = (\mu_{11}, \dots, \mu_{1p})'$ to have $\lfloor \kappa p^r \rfloor$ non-zero entries uniformly and randomly drawn from $\{1, \dots, p\}$, where κ is an integer. As before, we considered $r =$

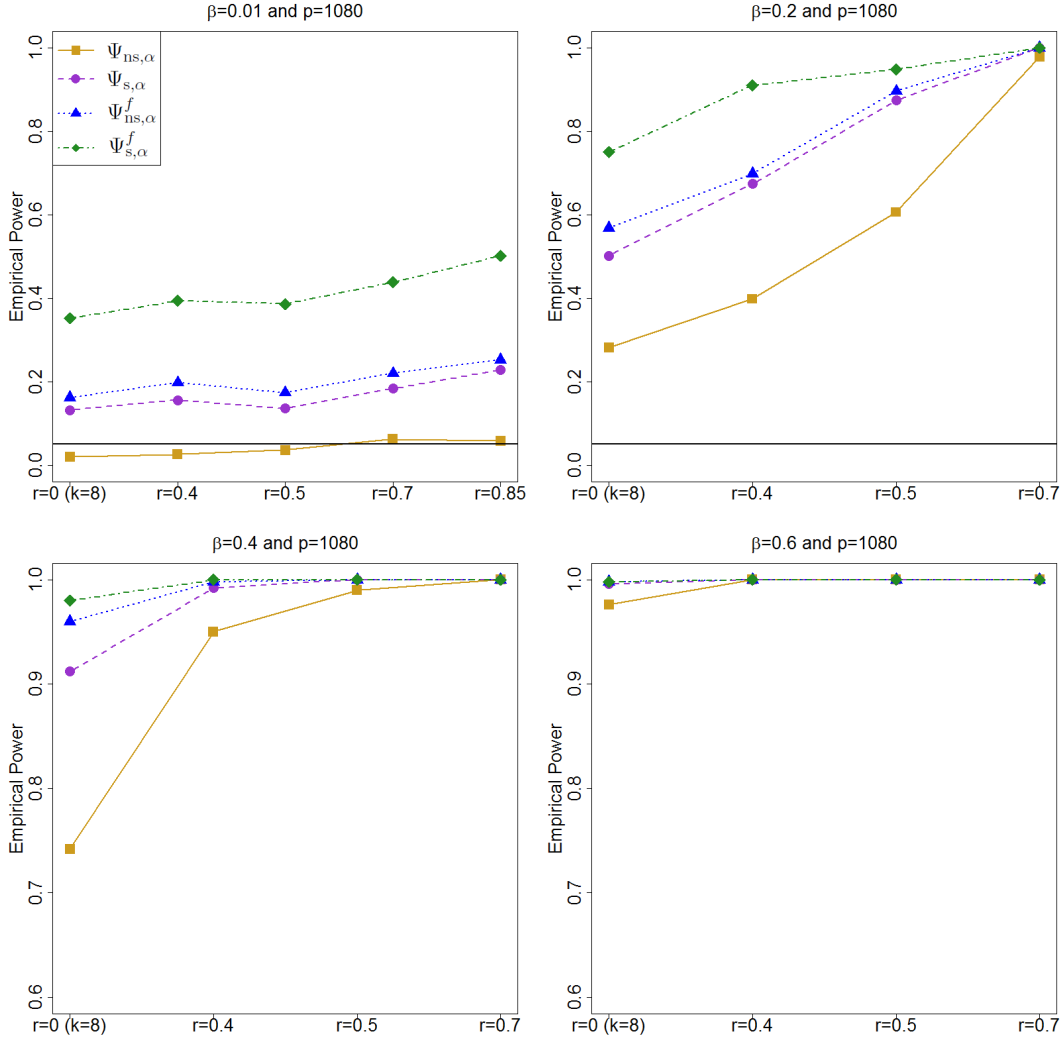


Figure 3: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of the signal strength (β) and sparsity ($1 - r$) for the one-sample problem (1.1) at 5% nominal significance for the Gaussian data and long range dependence covariance matrices in Model 3⁽¹⁾ when $n = 80$ and $p = 1080$.

0, 0.4, 0.5, 0.7 and 0.85, where $\kappa = 8$ if $r = 0$ and $\kappa = 1$ otherwise. The settings $r = 0$ and $r = 0.7$ or $r = 0.85$ mimic sparse and non-sparse scenarios, respectively. The magnitudes of non-zero entries $\mu_{1\ell}$ were set to be $\{2\beta\sigma_{\ell\ell}\log(p)(1/n + 1/m)\}^{1/2}$, where $\sigma_{\ell\ell}$ is the ℓ th diagonal entry of the pooled covariance matrix $\Sigma_{1,2}$ as in (2.5). We took $\beta = 0.01, 0.2, 0.4, 0.6$ and $\beta = 0.01$ to mimic the scenario of weak signals.

The following three models were used to generate random samples $\mathbf{X}_i = \mathbf{W}_i + \boldsymbol{\mu}_1$, $\mathbf{Y}_j = \mathbf{V}_j + \boldsymbol{\mu}_2$ for $i = 1, \dots, n$ and $j = 1, \dots, m$, where $\mathbf{W}_i = (W_{i,1}, \dots, W_{i,p})'$ and $\mathbf{V}_j = (V_{j,1}, \dots, V_{j,p})'$ are independent and identically distributed centered Gaussian random vectors with covariance

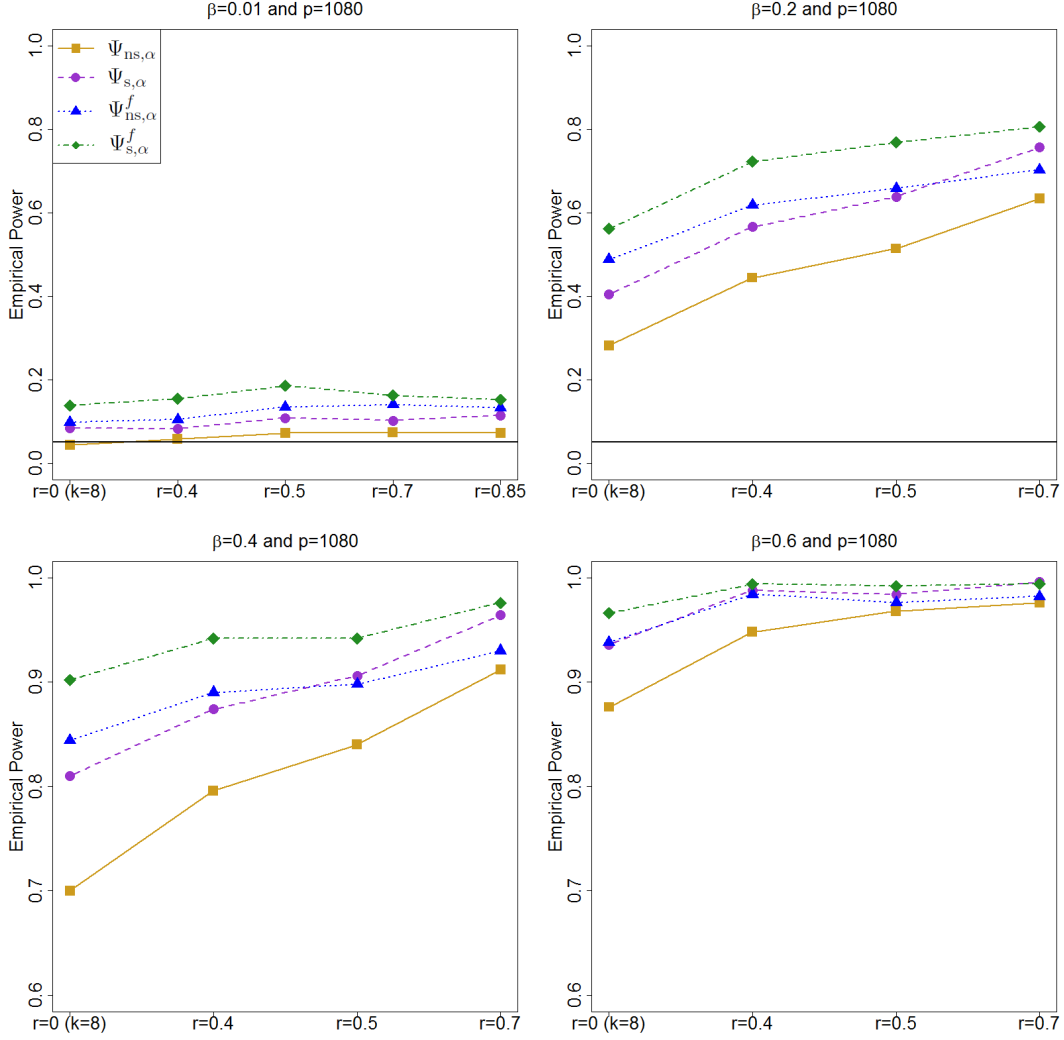


Figure 4: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of the signal strength (β) and sparsity ($1-r$) for the one-sample problem (1.1) at 5% nominal significance for the autoregressive process model, Model 4^(I), with t -distributed innovations when $n = 80$ and $p = 1080$.

matrices $\Sigma_1 = (\sigma_{1,k\ell})_{1 \leq k, \ell \leq p}$ and $\Sigma_2 = (\sigma_{2,k\ell})_{1 \leq k, \ell \leq p}$, respectively.

- (a) Model 1^(II) (Block diagonal covariance matrices): For $k = 1, \dots, p$ and $q = 1, 2$, $\sigma_{q,kk}$ are independent and identically drawn from $\text{Unif}(1, 2)$, $\sigma_{q,k\ell} = 0.7$ for $10(t-1)+1 \leq k \neq \ell \leq 10t$, where $t = 1, \dots, \lfloor p/10 \rfloor$, and $\sigma_{q,k\ell} = 0$ otherwise. This model was studied in [Cai, Liu and Xia \(2014\)](#).
- (b) Model 2^(II) (Non-sparse covariance matrices): Let $\mathbf{F} = (f_{k\ell})_{1 \leq k, \ell \leq p}$ with $f_{kk} = 1$, $f_{k,k+1} = f_{k+1,k} = 0.5$; $\mathbf{U}_q \sim \mathcal{U}(\mathcal{V}_{p,k_0})$, the uniform distribution on the Stiefel manifold for $q = 1, 2$;

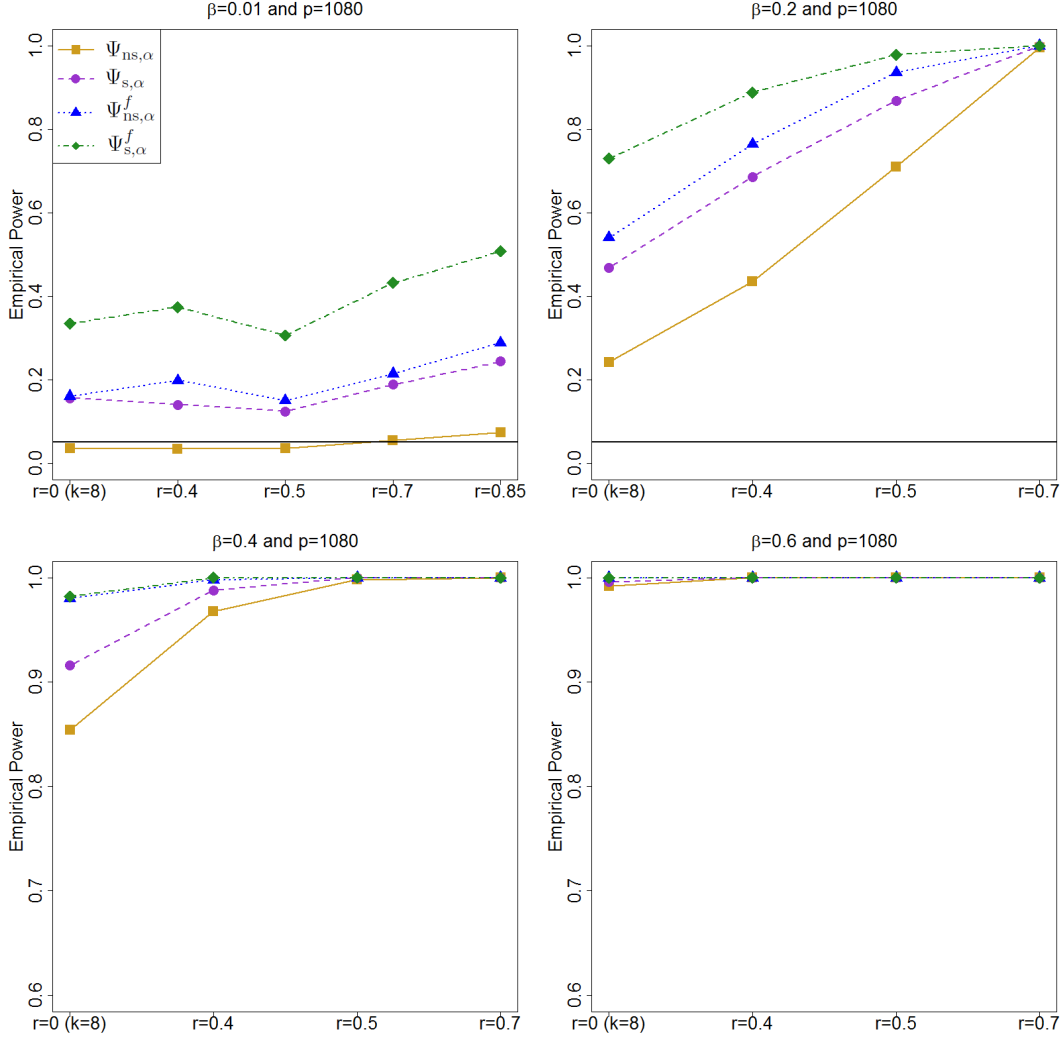


Figure 5: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of the signal strength (β) and sparsity ($1-r$) for the one-sample problem (1.1) at 5% nominal significance for the moving average process with Beta distributed innovations in Model 5^(I) when $n = 80$ and $p = 1080$.

and $\Theta = (\theta_{k\ell})_{1 \leq k, \ell \leq p}$ be a diagonal matrix with θ_{kk} independent and identically drawn from $\text{Unif}(1, 6)$. Set $k_0 = 10$ and put $\Sigma_q = \Theta^{1/2}(\mathbf{F} + \mathbf{U}_q \mathbf{U}_q') \Theta^{1/2}$ for $q = 1, 2$. A similar model was studied in Cai, Liu and Xia (2014).

- (c) Model 3^(II) (Long range dependence): Let $\theta_{11}, \dots, \theta_{1p}, \theta_{21}, \dots, \theta_{2p}$ be independent and identically drawn from $\text{Unif}(1, 2)$; for $q = 1, 2$, we took $\sigma_{q,kk} = \theta_{qk}$ and $\sigma_{q,k\ell} = \rho_\alpha(|k - \ell|)$ for $k \neq \ell$, where $\rho_\alpha(e) = \frac{1}{2}\{(e+1)^{2H} + (e-1)^{2H} - 2e^{2H}\}$ with $H = 0.9$.

Model 1^(II) imposes sparse covariance structures while Model 2^(II) and Model 3^(II) account for

non-sparse dependence structures. In addition, we considered the following two models with non-Gaussian data generation mechanisms to study the robustness of the proposed tests against Gaussianity.

- (d) Model 4^(II) (Autoregressive process of order one, AR(1), with t -distributed innovations): Generate independent and identically distributed p -variate random vectors $\mathbf{X}_i \sim t_{\omega_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{Y}_j \sim t_{\omega_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ for $i = 1, \dots, n$ and $j = 1, \dots, m$, where $t_{\omega_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $t_{\omega_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are the non-central multivariate t -distributions with non-central parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, degrees of freedoms $\omega_1 = 5, \omega_2 = 7$, and $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1 \leq k, \ell \leq p}$ with $\sigma_{1,k\ell} = 0.995^{|k-\ell|}$, $\boldsymbol{\Sigma}_2 = (\sigma_{2,k\ell})_{1 \leq k, \ell \leq p}$ with $\sigma_{2,k\ell} = 0.7^{|k-\ell|}$.
- (e) Model 5^(II) (Moving average process with Gamma distributed innovations): For $i = 1, \dots, n$, $j = 1, \dots, m$, and $k = 1, \dots, p$, set $X_{ik} = \rho_{1,1}Z_{i,k} + \rho_{1,2}Z_{i,k+1} + \dots + \rho_{1,p}Z_{i,k+p-1} + \mu_{1k}$ and $Y_{jk} = \rho_{2,1}\tilde{Z}_{j,k} + \rho_{2,2}\tilde{Z}_{j,k+1} + \dots + \rho_{2,p}\tilde{Z}_{j,k+p-1} + \mu_{2k}$, where $\rho_{1,\ell}$ are independent and identically drawn from $0.6\text{Unif}(-1, 1) + 0.4\delta_0$ and $\rho_{2,\ell}$ are independent and identically drawn from $0.8\text{Unif}(-1, 1) + 0.2\delta_0$ for $\ell = 1, \dots, p$, where δ_0 is the point mass at 0, $\{Z_{i,k}\}$ and $\{\tilde{Z}_{j,k}\}$ are independent random variables with a common centered Gamma(1, 4) and Gamma(4, 1) distributions, respectively.

Model 4^(I) and Model 5^(I) impose non-sparse covariance structures. Simulation results show that the proposed tests are valid in these non-sparse models with non-Gaussian sampling distributions.

In each model, we generated data with sample sizes $n = m = 40$ and $n = m = 80$. The dimension p took values 120, 360 and 1080. The empirical power and level of significance were computed based on 1500 simulations.

The numerical results on the proposed tests $\Psi_{\text{ns},\alpha}$, $\Psi_{\text{s},\alpha}$, $\Psi_{\text{ns},\alpha}^f$ and $\Psi_{\text{s},\alpha}^f$ and the HC2, CQ and CLX tests are summarized in Table 2 and Figures 6-10. Table 2 displays the empirical sizes. It can be seen that in all the models, the empirical size for the proposed non-studentized tests $\Psi_{\text{ns},\alpha}$ and $\Psi_{\text{ns},\alpha}^f$ are reasonably close to the nominal level 0.05 for both $n = m = 40$ and $n = m = 80$. Comparing the results for $n = m = 40$ and $n = m = 80$, the studentized tests, $\Psi_{\text{s},\alpha}$ and $\Psi_{\text{s},\alpha}^f$, have slightly inflated significance when the sample size is relatively small but improve when the sample size increases. Additionally, the CLX test fails to maintain the nominal size for Model 4^(II) due to the strong dependency in the covariance structures. Analogous to the observation in Section 4.1, it is difficult for the HC2 procedure to maintain the nominal significance when the sample size is small or the dependency is strong and complex. The CQ test maintains the nominal significance reasonably well in all the models.

To evaluate the power, we compared the proposed tests with the CQ and CLX tests for $n = m = 80$ and $p = 1080$. It can be seen that the tests with screening, $\Psi_{\text{ns},\alpha}^f$ and $\Psi_{\text{s},\alpha}^f$, outperform both the CQ and CLX tests against alternatives with sparse signals ($r = 0$) for different signal strength β . On the other hand, all the tests perform similarly when the signals become less sparse

Model 1 ^(II)		Model 2 ^(II)				Model 3 ^(II)				Model 4 ^(II)				Model 5 ^(II)			
tests / p	120	360	1080	120	360	1080	120	360	1080	120	360	1080	120	360	1080	120	360
$(n, m) = (40, 40)$																	
$\Psi_{\text{ns},\alpha}$	0.039	0.041	0.041	0.042	0.044	0.039	0.051	0.036	0.030	0.052	0.036	0.042	0.034	0.030	0.032	0.034	0.032
$\Psi_{\text{s},\alpha}$	0.094	0.112	0.125	0.092	0.097	0.116	0.096	0.095	0.130	0.086	0.090	0.092	0.094	0.141	0.158	0.094	0.141
$\Psi_{\text{ns},\alpha}^f$	0.055	0.048	0.057	0.049	0.055	0.054	0.063	0.045	0.044	0.055	0.039	0.052	0.044	0.038	0.045	0.044	0.038
$\Psi_{\text{s},\alpha}^f$	0.092	0.120	0.152	0.098	0.131	0.053	0.092	0.093	0.160	0.090	0.094	0.094	0.092	0.174	0.180	0.092	0.174
HC2	0.140	0.202	0.343	0.127	0.156	0.308	0.138	0.221	0.315	0.280	0.344	0.391	0.155	0.210	0.447	0.155	0.210
CQ	0.044	0.049	0.034	0.046	0.049	0.051	0.046	0.059	0.051	0.064	0.066	0.054	0.046	0.047	0.051	0.046	0.047
CLX	0.101	0.103	0.138	0.081	0.087	0.098	0.085	0.093	0.116	0.204	0.181	0.137	0.081	0.127	0.157	0.081	0.127
$(n, m) = (80, 80)$																	
$\Psi_{\text{ns},\alpha}$	0.054	0.039	0.046	0.053	0.040	0.040	0.042	0.038	0.049	0.046	0.045	0.047	0.041	0.036	0.035	0.041	0.036
$\Psi_{\text{s},\alpha}$	0.074	0.062	0.086	0.058	0.064	0.090	0.064	0.071	0.090	0.059	0.065	0.074	0.070	0.090	0.097	0.070	0.090
$\Psi_{\text{ns},\alpha}^f$	0.065	0.052	0.060	0.063	0.050	0.058	0.052	0.044	0.063	0.047	0.048	0.056	0.046	0.056	0.045	0.046	0.056
$\Psi_{\text{s},\alpha}^f$	0.088	0.076	0.098	0.070	0.080	0.093	0.073	0.084	0.093	0.062	0.069	0.086	0.086	0.092	0.099	0.086	0.092
HC2	0.112	0.132	0.170	0.079	0.076	0.133	0.092	0.147	0.186	0.237	0.292	0.377	0.099	0.150	0.207	0.377	0.099
CQ	0.046	0.039	0.048	0.048	0.038	0.048	0.038	0.044	0.049	0.044	0.054	0.056	0.051	0.053	0.053	0.054	0.053
CLX	0.107	0.090	0.104	0.057	0.057	0.089	0.058	0.060	0.083	0.289	0.352	0.297	0.059	0.087	0.098	0.297	0.059

Table 2: Empirical sizes of the proposed tests (non-studentized without screening $\Psi_{\text{ns},\alpha}$, studentized without screening $\Psi_{\text{s},\alpha}$, non-studentized with screening $\Psi_{\text{ns},\alpha}^f$, and studentized with screening $\Psi_{\text{s},\alpha}^f$) for the two-sample problem (1.2), along with those of the tests by [Delaigle, Hall and Jin \(2011\)](#) (HC2), [Chen and Qin \(2010\)](#) (CQ), and [Cai, Liu and Xia \(2014\)](#) (CLX) at 5% nominal significance. Models with Gaussian data and block diagonal, non-sparse or long range dependence covariance matrices, the autoregressive process model with t -distributed innovations, and the moving average processes model with Gamma distributed innovations are considered when $n = m = 40, 80$ and $p = 120, 360, 1080$.

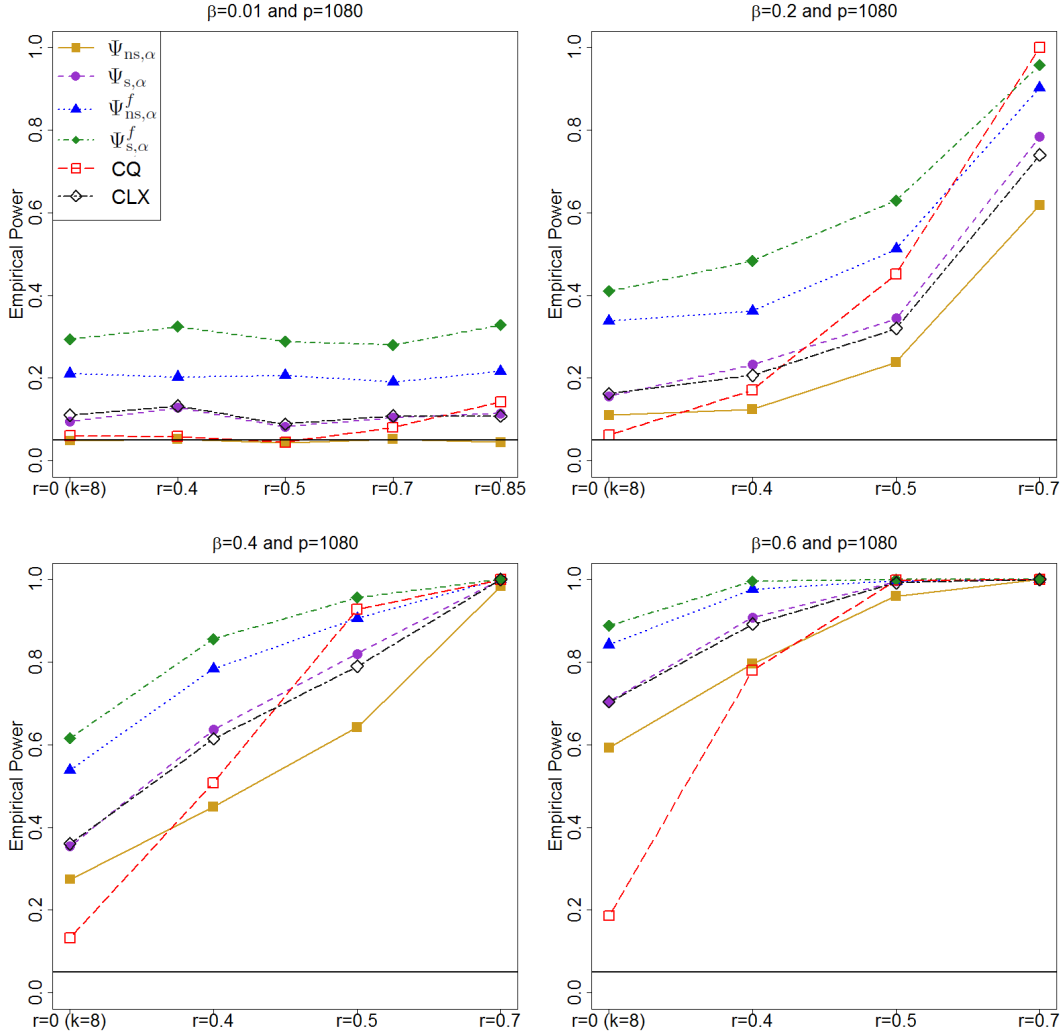


Figure 6: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of signal strength (β) and sparsity ($1-r$) for the two-sample problem (1.2), along of those of the tests by Chen and Qin (2010) (CQ) and Cai, Liu and Xia (2014) (CLX) at 5% nominal significance for the Gaussian data and block diagonal covariance matrices in Model 1^(II) when $n = m = 80$ and $p = 1080$.

and strong (see the right bottom panels in Figures 6 to 8 and Figure 10). The CQ test gains more powers when signals become less sparse, as expected for sum of squares-type statistics. Its power approaches to those of the proposed tests with screening $\Psi_{ns,\alpha}^f$ and $\Psi_{s,\alpha}^f$ when the signals become less sparse and stronger ($r \geq 0.5, \beta \geq 0.4$) in all the models except Model 4^(II). In Model 4^(II), all the proposed tests outperform the CQ test substantially (Figure 9) as the sum of squares-type test statistics may lose power for heavy tailed sampling distributions. The CLX test performs similarly to the proposed tests without screening $\Psi_{ns,\alpha}$ and $\Psi_{s,\alpha}$, but is outperformed by the proposed tests

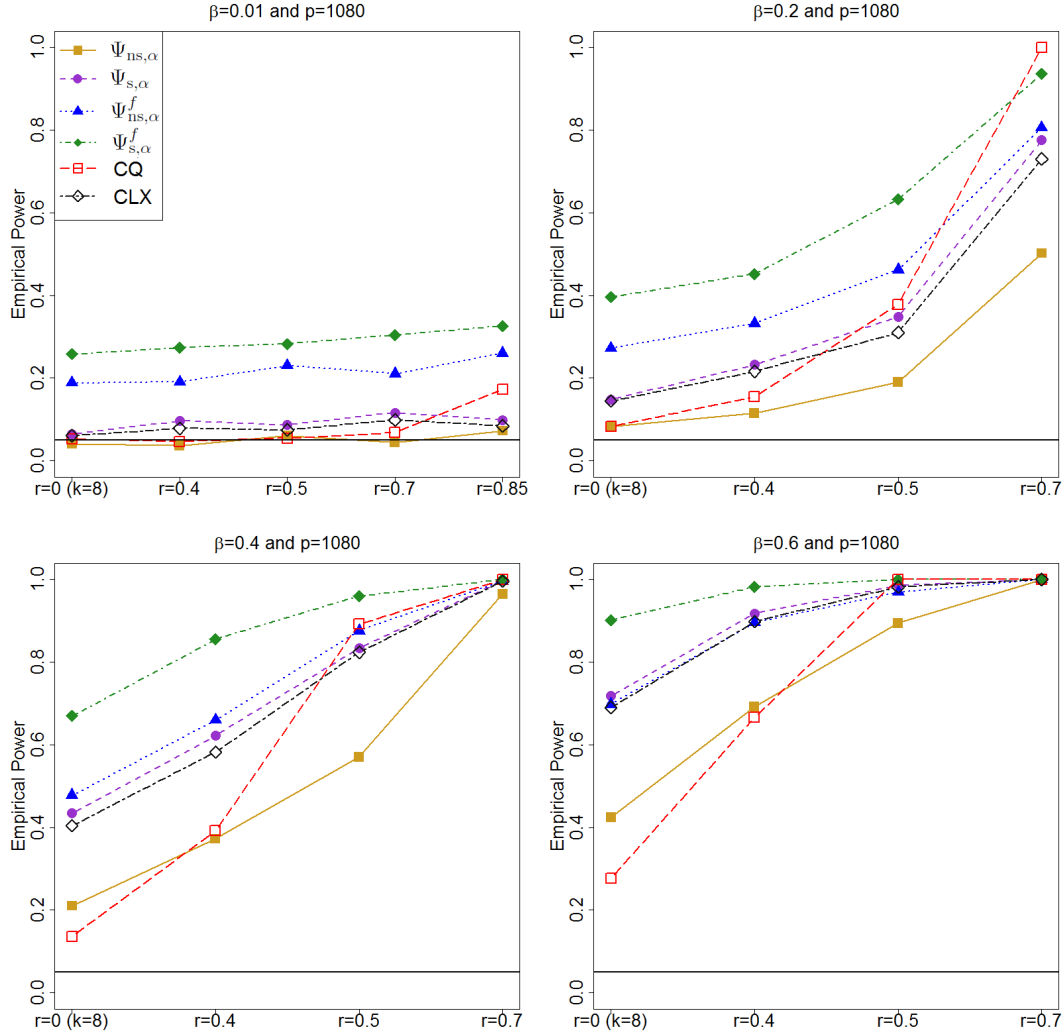


Figure 7: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of signal strength (β) and sparsity ($1-r$) for the two-sample problem (1.2), along of those of the tests by Chen and Qin (2010) (CQ) and Cai, Liu and Xia (2014) (CLX) at 5% nominal significance for the Gaussian data and non-sparse covariance matrices in Model 2^(II) when $n = m = 80$ and $p = 1080$.

with screening for all settings. The simulation results agree with the heuristic discussion and the theoretical justification that the screening step substantially improves the power of proposed tests. Similar to the observations in Section 2.1.1, the non-studentized test with screening $\Psi_{ns,\alpha}^f$ is preferable in practice whenever the sample size is relatively small. More extensive simulations were carried out for dimensions $p = 120$ and 360 , from which the comparisons are consistent with the cases that are reported here. The empirical powers of all the tests also increase in p . The additional simulation results are placed in the online supplementary material.

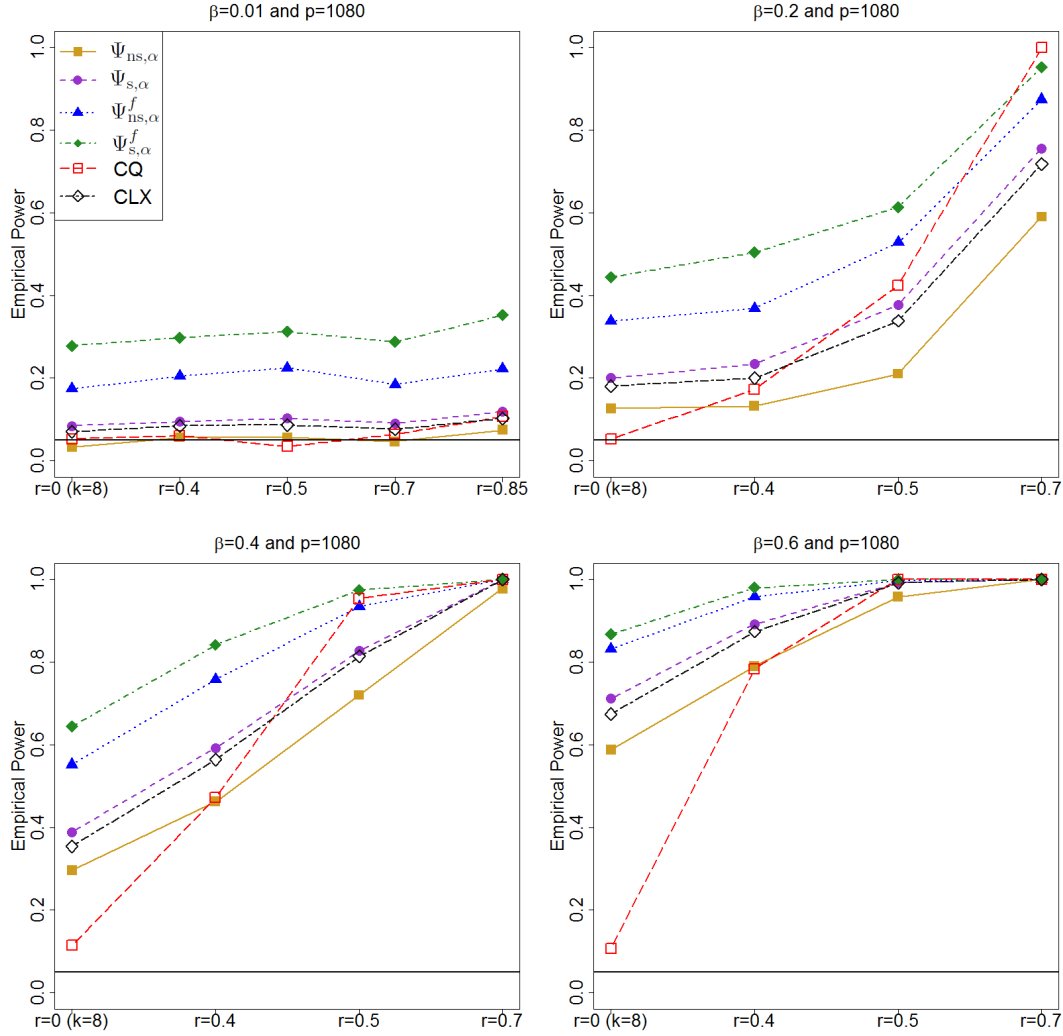


Figure 8: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of signal strength (β) and sparsity ($1 - r$) for the two-sample problem (1.2), along of those of the tests by [Chen and Qin \(2010\)](#) (CQ) and [Cai, Liu and Xia \(2014\)](#) (CLX) at 5% nominal significance for the Gaussian data and long range dependence covariance matrices in Model 3^(II) when $n = m = 80$ and $p = 1080$.

To sum up, the numerical results show that the proposed tests, particularly the studentized tests and the non-studentized test with screening, $\Psi_{s,\alpha}$, $\Psi_{s,\alpha}^f$ and $\Psi_{ns,\alpha}^f$, outperform the existing methods when the covariance structure is non-sparse and complex. The proposed tests are robust against both unknown covariance structures and Gaussianity. The non-studentized test with screening $\Psi_{ns,\alpha}^f$ maintains the nominal significance for small sample sizes and has good powers against sparse alternatives, which is recommended for practical applications whenever the sample sizes are limited.

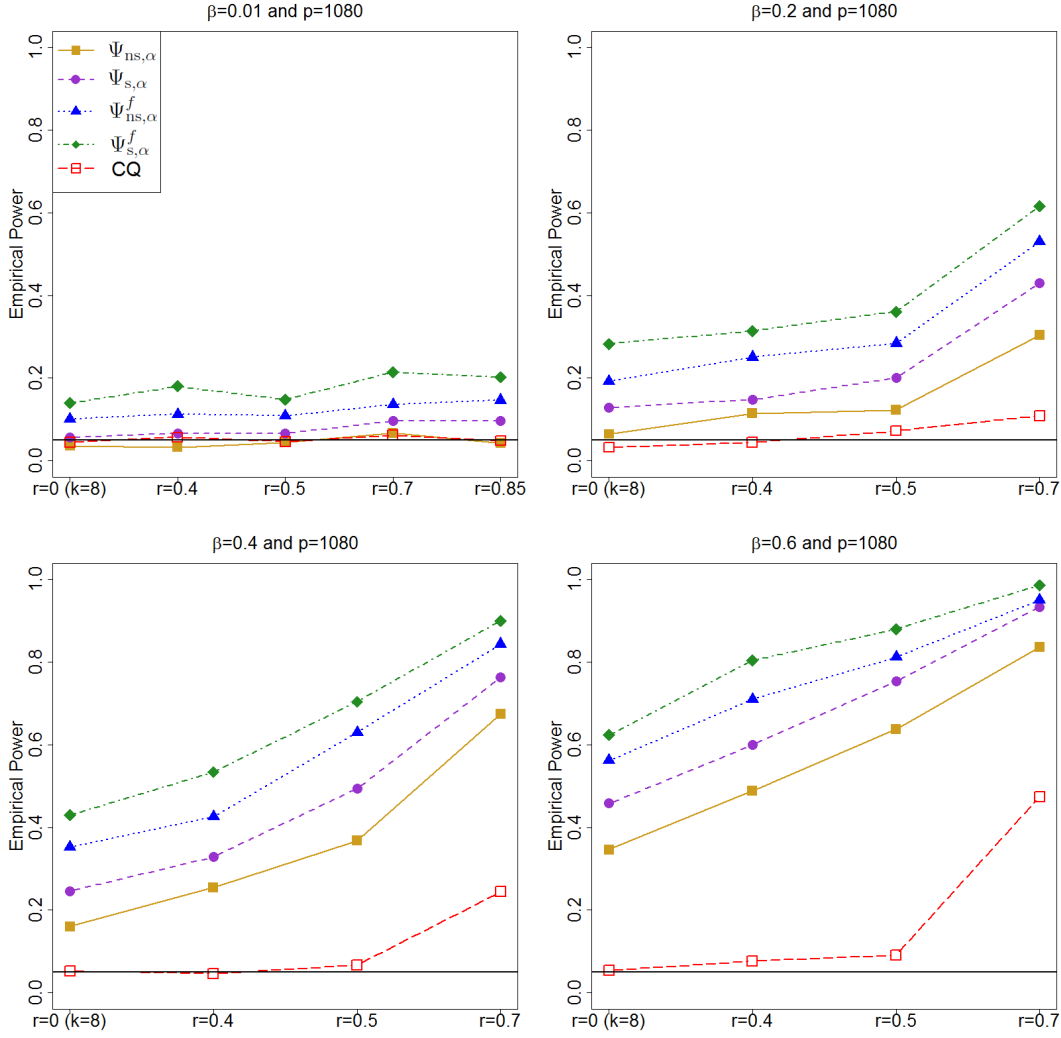


Figure 9: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of signal strength (β) and sparsity ($1-r$) for the two-sample problem (1.2), along of those of the test by [Chen and Qin \(2010\)](#) (CQ) at 5% nominal significance for the autoregressive process model, Model 4^(II), with t -distributed innovations when $n = m = 80$ and $p = 1080$.

5 Empirical study

To achieve certain biological functions, genes tend to work in groups which are known as gene-sets. Analysis and interpretation based on gene-sets derive more power than focusing on individual gene in extracting biological insights ([Subramanian et al., 2005](#)). It has drawn increasing attentions to identify gene-sets associated with biological states of interest, such as gene-sets associated with disease developments or evolutionary mutations, in biomedical research and general genome

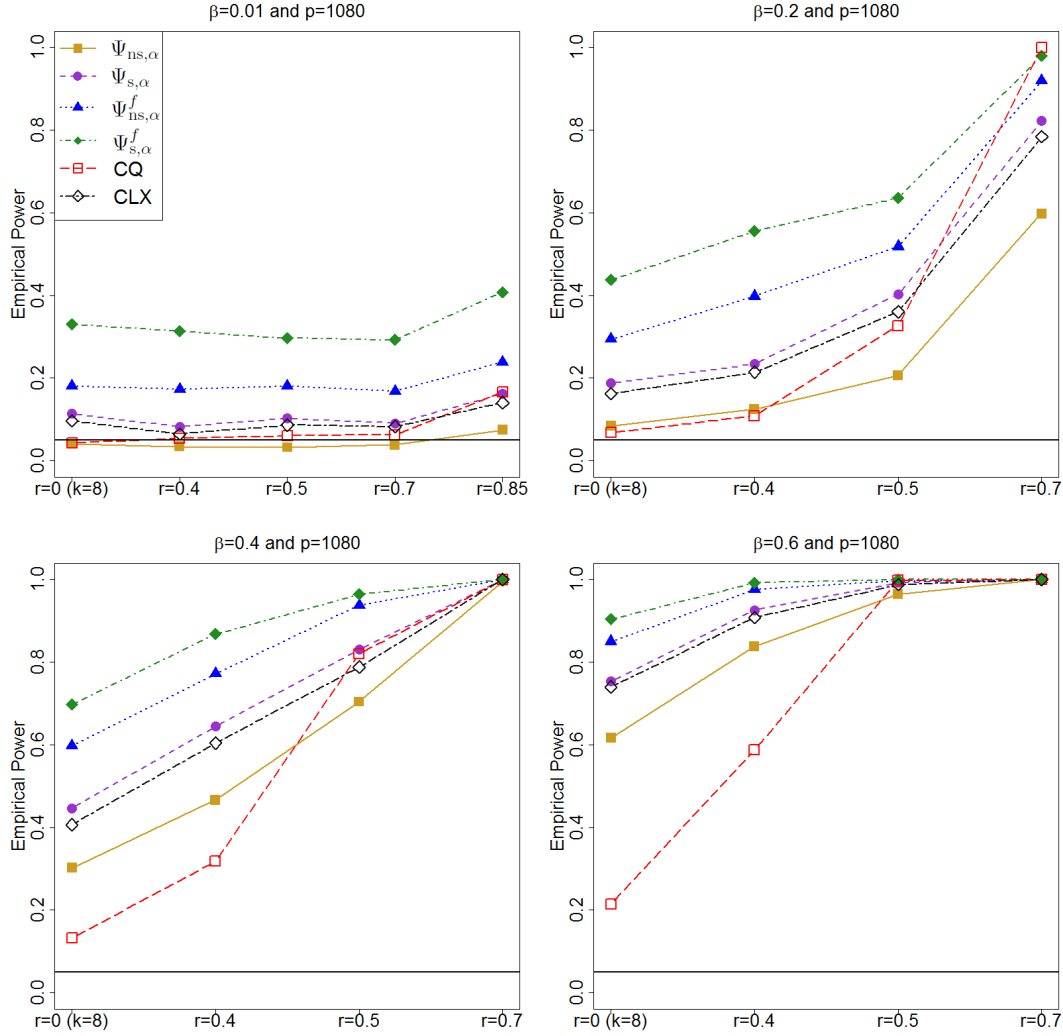


Figure 10: Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi_{ns,\alpha}^f$, and studentized with screening $\Psi_{s,\alpha}^f$) against alternatives with different levels of signal strength (β) and sparsity ($1-r$) for the two-sample problem (1.2), along of those of the tests by [Chen and Qin \(2010\)](#) (CQ) and [Cai, Liu and Xia \(2014\)](#) (CLX) at 5% nominal significance for the moving average process with Gamma distributed innovations in Model 5^(II) when $n = m = 80$ and $p = 1080$.

analysis ([Subramanian et al., 2005](#); [Efron and Tibshirani, 2007](#); [Recknor, Nettleton and Reecy, 2008](#); [Thomas, Joshi and Klaperb, 2011](#)). The gene-sets are technically defined in gene ontology (GO) system which provides structured vocabularies producing the names of the sets or known as GO terms. Three categories of gene ontologies of interest have been identified: biological processes (BP), cellular components (CC) and molecular functions (MF). That is, a particular gene-set will belong to one of these three groups of GO terms.

Statistically, identifying interesting gene-sets out of G candidate gene-sets $\mathcal{S}_1, \dots, \mathcal{S}_G$ based

on independent samples from two biological states is equivalent to test hypotheses

$$H_{0s} : \boldsymbol{\mu}_{1,s} = \boldsymbol{\mu}_{2,s} \text{ versus } H_{1s} : \boldsymbol{\mu}_{1,s} \neq \boldsymbol{\mu}_{2,s}$$

for $s = 1, \dots, G$, where the p_s dimensional vector $\boldsymbol{\mu}_{i,s}$ models the mean expression levels of p_s genes in the gene-set \mathcal{S}_s under the biological condition i for $i = 1, 2$, which may correspond to control and mutation groups or health and diseased groups in practice. It is common that gene-sets overlap with each other as one particular gene may belong to several functional groups, and the size of a gene-set p_s usually range from a small to a very large number. The selection of gene-sets therefore encounters both multiplicity and high dimensionality. As suggested by [Chen and Qin \(2010\)](#), we applied the proposed tests to each gene-set. With p -values obtained for all G gene-sets, similar to [Chen and Qin \(2010\)](#), we further employed multiple testing procedures such as the Bonferroni method for controlling the family-wise error rate or procedures such as the Benjamini-Hochberg procedure ([Benjamini and Hochberg, 1995](#)) for controlling the false discovery rate (FDR) to identify significant gene-sets. In practice, we compute the empirical p -values of the proposed tests by $\widehat{p_v} = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{ns,\ell}|_\infty > T_{ns}^{(II)}\}$ or $\widehat{p_v} = M^{-1} \sum_{\ell=1}^M I\{|\mathbf{W}_{s,\ell}|_\infty > T_s^{(II)}\}$. Here $\{\mathbf{W}_{ns,1}, \dots, \mathbf{W}_{ns,M}\}$ and $\{\mathbf{W}_{s,1}, \dots, \mathbf{W}_{s,M}\}$ are two random samples independently generated from $N(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{1,2})$ and $N(\mathbf{0}, \widehat{\mathbf{R}}_{1,2})$, where $\widehat{\boldsymbol{\Sigma}}_{1,2}$ and $\widehat{\mathbf{R}}_{1,2}$ are defined via replacing $\widetilde{\boldsymbol{\Sigma}}_1$ and $\widetilde{\boldsymbol{\Sigma}}_2$ in (2.5) by sample covariance matrices $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$, respectively. The empirical p -values for the proposed two-step procedure are computed similarly.

We applied the above procedure to a human acute lymphoblastic leukemia (ALL) dataset which is available at <http://www.ncbi.nlm.nih.gov>. The data contains gene expression levels from microarray experiments for patients suffering from ALL of either T-lymphocyte type or B-lymphocyte type leukemia. This dataset was originally analyzed by [Chiaretti et al. \(2004\)](#) to provide insight into the genetic mechanism on ALL development. The data was also analyzed by [Dudoit, Keles and van der Laan \(2011\)](#) and [Chen and Qin \(2010\)](#), among others using different methodologies. To illustrate the proposed tests, we focus on the 75 patients of B-lymphocyte type leukemia, who were classified into two groups: 35 patients with BCR/ABL fusion and 40 patients with cytogenetically normal NEG, i.e. $n = 35$ and $m = 40$. We employed the approach in [Gentleman et al. \(2005\)](#) to conduct preliminary data processing. To focus on high dimensional scenarios, we also excluded gene-sets with small numbers of genes that only gene-sets with $p_s \geq 19$ were retained. It remained $G = 1853, 262$ and 284 unique GO terms in the BP, CC and MF categories, respectively. The largest gene-set contained $p_s = 3050$ genes in the BP category, $p_s = 3145$ genes in the CC category, and $p_s = 3040$ genes in the MF category. Identifications of gene-sets associated to the BCR/ABL fusion display biological insights on the development of B-lymphocyte type leukemia and provide lists of functional groups for potential clinical treatments. Our aim is therefore to identify gene-sets with significantly different expression levels between the BCR/ABL and NEG groups for each of the three categories.

The sample size of the ALL data is relatively small comparing to the maximum p_s , we therefore employed the proposed two-sample non-studentized tests $\Psi_{ns,\alpha}$ and $\Psi_{ns,\alpha}^f$ in analysis as suggested by simulation studies in Section 4. Based on empirical p -values, we further employed the Benjamini-Hochberg (BH) procedure for controlling the FDR at 0.015 and identify significant gene-sets. For the proposed tests, we let $M = 50000$ and used the sample covariance matrices to generate samples. Simulation studies in Section 4 have shown that the test by Cai, Liu and Xia (2014) may inflate type I error rate for small sample size, we therefore only consider the test by Chen and Qin (2010) (CQ) as a comparison. For each category, the numbers of gene-sets being identified are summarized in Table 3. The proposed two-stage test, $\Psi_{ns,\alpha}^f$, identified more gene-sets than other methods. Particularly, $\Psi_{ns,\alpha}^f$ found more disease associated gene-sets than $\Psi_{ns,\alpha}$ which reflects the power improvement of the proposed two-stage testing procedure as discussed before. The overlap among gene-sets increases in the number of genes within each gene-set p_s (Chen and Qin, 2010), which explains the relatively large number of identified disease-associated gene-sets in Table 3.

GO Category	Total	$\Psi_{ns,\alpha}$	$\Psi_{ns,\alpha}^f$ and CQ			$\max_s p_s$	$\min_s p_s$	$[\bar{p}_s]$
			$\Psi_{ns,\alpha}^f$ only	Both	CQ only			
BP	1853	1082	302	1065	136	3050	20	150
CC	262	93	33	117	28	3145	19	280
MF	284	136	63	133	14	3040	19	157

Table 3: Numbers of identified BCR/ABL associated gene-sets for each GO category using different tests in conjunction with the BH procedure for controlling FDR at 0.015. Columns labeled by the name of tests records the number of identified gene-sets by the corresponding testing procedures, where $\Psi_{ns,\alpha}$ and $\Psi_{ns,\alpha}^f$ are the proposed non-studentized tests without and with screening, and CQ stands for the test by Chen and Qin (2010).

By carefully investigating the gene-sets identified by both the proposed tests $\Psi_{ns,\alpha}$ and $\Psi_{ns,\alpha}^f$, we found that gene-sets GO:0005758 (mitochondrial intermembrane space) and GO:0004860 (protein kinase inhibitor activity) were identified as diseases-associated in the CC and MF categories. The functions of these two interesting gene-sets were recently studied and recognized associated with the development of ALL (Brinkmann and Kashkar, 2014; Cui et al., 2009). Particularly, the protein kinase inhibition has been considered to be essential for the mechanism of T-lymphocyte type ALL (Cui et al., 2009) and our finding suggests its connection with B-lymphocyte type ALL as well. In the supplementary materials, we list the top 15 gene-sets that were identified to be diseases-associated by the proposed non-studentized test with screening $\Psi_{ns,\alpha}^f$ but missed by the CQ test with the FDR controlled at 0.015. The association of these gene-sets with the ALL may deserve further biological validations using the polymerase chain reaction.

6 Discussion

In this paper, we study hypothesis testing of high dimensional mean vectors in both the one-sample and two-sample settings. Unlike the existing tests, the proposed procedures enforce no structural assumptions on the unknown covariance matrices, and are powerful against sparse alternatives.

The test statistics are taken as the extreme values of partial sums or self-normalized partial sums, whose distributions can be approximated by the extreme values of centred Gaussian vectors with the same covariance structure as the p -vector of all marginal statistics. The uniform convergence of the approximated distributions to the target distributions of the test statistics is guaranteed whenever $\|\widehat{\Sigma} - \Sigma\|_\infty + \|\widehat{\mathbf{R}} - \mathbf{R}\|_\infty = o_P(1)$; that is, the estimates of the covariance and correlation matrices are max-norm consistent, which can be easily met using the sample covariance matrices under mild moment conditions. Therefore, the proposed tests allow for very general form of the unknown covariance matrices even for non-invertible covariances that model perfect correlated variables. This important feature of the proposed tests is supported by simulation studies, whereas most existing methods may encounter difficulties in maintaining the nominal sizes for complex and/or strong dependency. Additional numerical experiments regarding perfectly correlated variables are also included in the supplementary material.

To further improve the power of proposed tests, we developed a preliminary feature screening step to construct the two-step testing procedures $\Psi_{\text{ns},\alpha}^f$ and $\Psi_{\text{s},\alpha}^f$. In principle, the pre-screening is based on marginal signal-to-noise ratios and improves the numerical performance of the proposed tests against sparse alternatives. Theoretical justifications of the two-step tests are provided by Theorems 5 and 6. Numerical studies demonstrate their superior performance over the others. For example, when the sample size is relatively small, the non-studentized test with screening $\Psi_{\text{ns},\alpha}^f$ is more robust than others yet maintains satisfactory powers. Particularly, the data analysis for gene-set selections shows that the two-step testing procedures are capable of identifying diseases-associated gene-sets that were missed by the other peer tests. It is appealing to notice the connection between the preliminary feature screening and augment classifications in high dimensional settings (Fan and Fan, 2008). For a binary classification problem, which is essentially a two-sample problem, the preliminary feature screening step might be employed, in conjunction with a broader class of statistics such as the Kolmogorov-Smirnov test, to select features with discriminant powers and therefore improve the prediction performance.

Supplementary Material

Supplementary material available online includes technical proofs, derivation of the main theorems, more extensive simulation results, and details regarding real data analysis.

References

- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley-Interscience, New York.
- Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.
- Brinkmann, K. and Kashkar, H. (2014). Targeting the mitochondrial apoptotic pathway: a preferred approach in hematologic malignancies? *Cell Death and Disease*, **5**, e1098.
- Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199–227.
- Bickel, P. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577–2604.
- Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* **41**, 2786–2819.
- Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, **106**, 672–684.
- Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B*, **76**, 349–372.
- Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, **40**, 2014–2042.
- Cai, T. T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, **40**, 2389–2420.
- Castagna, J. P., Sun, S. and Siegfried, R. W. (2003). Instantaneous spectral analysis: detection of low-frequency shadows associated with hydrocarbons. *The Leading Edge*, **22**, 120–127.
- Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, **41**, 2123–2148.
- Chen, S. X. and Qin, Y. (2010). A two sample test for high dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808–835.
- Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

- Cui, J., Wang, Q., Wang, J., Lv, M., Zhu, N., Li, Y., Feng, J., Shen, B. and Zhang J. (2009). Basal c-Jun NH2-terminal protein kinase activity is essential for survival and proliferation of T-cell acute lymphoblastic leukemia cells. *Molecular Cancer Therapeutics*, **8**, 3214–3222.
- Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student’s t -statistic. *Journal of the Royal Statistical Society, Series B*, **73**, 283–301.
- Dembo, A. and Shao, Q.-M. (2006). Large and moderate deviations for Hotelling’s T^2 -statistic. *Electronic Communications in Probability*, **11** 149–159.
- Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, **32**, 962–994.
- Dudoit, S., Keles, S. and van der Laan, M. J. (2008). Multiple tests of associations with biological annotation metadata. *Institute of Mathematical Statistics. Collections*, **2**, 153–218.
- Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1**, 107–129.
- Fan, J. and Fan, Y. (2008). High-dimensional classification using feature annealed independence rules. *The Annals of Statistics*, **36**, 2605–2637.
- Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S. and Huber, W. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, New York.
- James, D., Clymer, B. D. and Schmalbrock, P. (2001). Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging*, **13**, 876–881.
- Katsani, K. R., Irimia, M., Karapiperis, C., Scouras, Z. G., Blencowe, B. J., Promponas, V. J. and Ouzounis, C. A. (2014). Functional genomics evidence unearths new moonlighting roles of outer ring coat nucleoporins. *Scientific Reports*, **4**, 4655.
- Liu, W. and Shao, Q.-M. (2013). A Cramér moderate deviation theorem for Hotelling’s T^2 -statistic with applications to global tests. *The Annals of Statistics*, **41**, 296–322.
- Liu, W., Lin, Z. Y. and Shao, Q.-M. (2008). The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability*, **18**, 2337–2366.
- Martens, J. W. M. *et al.* (2005). Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer Research*, **65**, 4101–4117.
- Recknor, J., Nettleton, D. and Reecy, J. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

- Srivastava, M. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, **100**, 518–532.
- Srivastava, M. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, **99**, 386–402.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Science*, **102**, 15545–15550.
- Thomas, M. A., Joshi, P. P. and Klaperb, R. D. (2011). Gene-class analysis of expression patterns induced by psychoactive pharmaceutical exposure in fathead minnow (*Pimephales promelas*) indicates induction of neuronal systems. *Comparative Biochemistry and Physiology C*, **155**, 109–120.
- van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.
- Wolen, A. R. and Miles, M. F. (2012). Identifying gene networks underlying the neurobiology of ethanol and alcoholism. *Alcohol Research: Current Reviews*, **34**, 306–317.
- Zhong, P.-S., Chen, S. X. and Xu, M. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics*, **41**, 2820–2851.